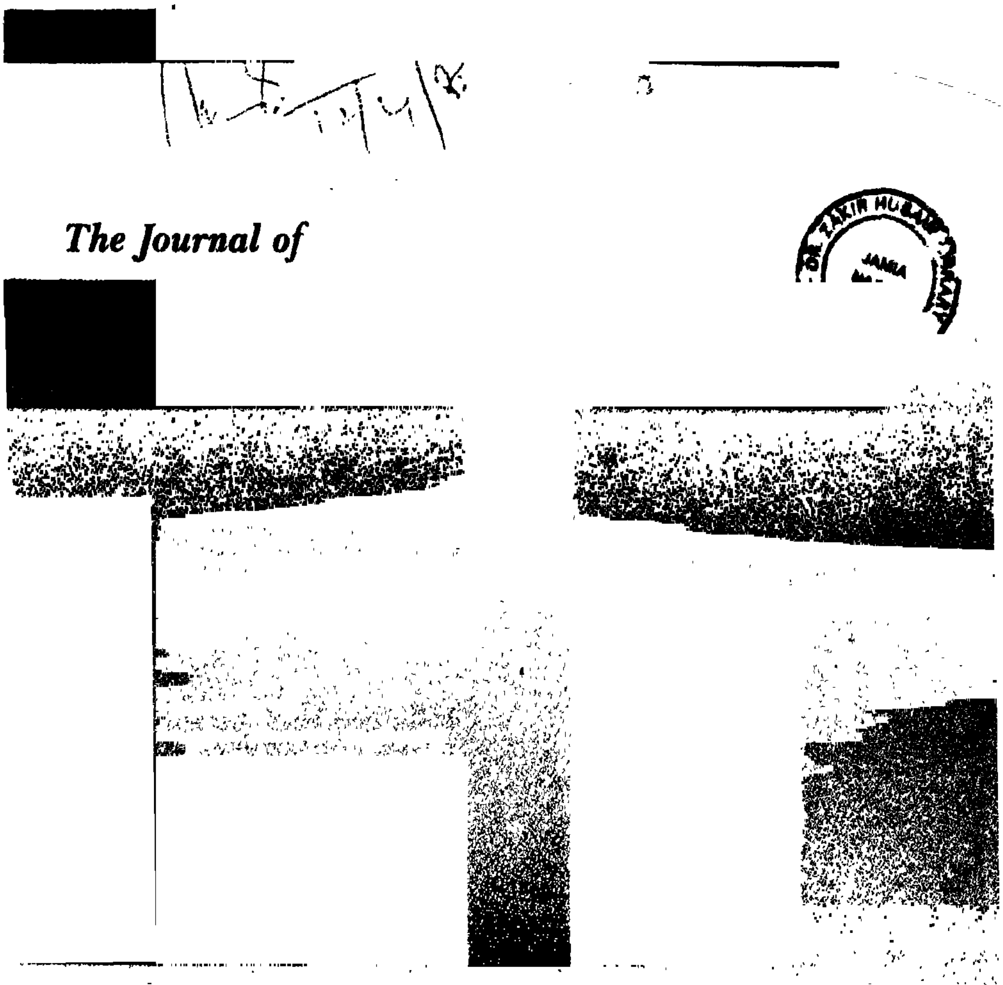
The Journal of



DR. ZAKIR HUSAIN LIBRARY
JAMIA

11

Winter 1988

Velma Montoya Thompson: associate professor of finance, Faculty of Finance, Real Estate and Law, California State University Polytechnic, Pomona, September 1988.

Randall M. Wigle: assistant professor, Wilfrid Laurier University, July 1, 1988.

Martin Zelder: instructor, economics department, College of William and Mary, August 1988.

Leaves for Special Appointments

James Hess, North Carolina State University: visiting professor, German Research Association Chair of Organized Theory, Koblenz School of Corporate Management, Koblenz, Germany, August 16, 1988–May 15, 1989.

David N. Hyman, North Carolina State University: guest scholar, Brookings Institution, Washington, D.C., August 16, 1988–December 31, 1989.

Dennis W. Jansen: Visiting Scholar at The Federal Reserve Bank of St. Louis, September 1, 1988–September 1, 1989.

Resignations

James R. Barth, Department of Economics, George Washington University; Federal Home Loan Bank Board.

Simón Teitel, The Catholic University of America, January 11, 1988.

Robert K. Triest: assistant professor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Miscellaneous

Zvi Griliches, Professor of Economics at Harvard University, is the winner of the W. S. Woytinski Lectureship Award for 1989. This award, given in recognition of significant research contributions in economics, commemorates W. S. Woytinski whose career was devoted to empirical research on human resource policy issues.

Professor Baldev Raj will visit the Institute of Economics, Norwegian School of Economics and Business Administration, Bergen, Norway, June 15, 1988, to August 31, 1988, funded by a fellowship from the Norwegian Research Council for Science and the Humanities.

To Departmental Representatives and Executive Officers

When submitting information for inclusion in "Notes," please observe the following guidelines:

Calls for papers, notice of professional meetings, and other announcements of interest to economists should be submitted in one paragraph that contains all relevant information.

News of individual members should be labeled as to category: (1) deaths; (2) retirements; (3) foreign scholars visiting the US or Canada; (4) promotions; (5) administrative appointments; (6) new appointments; (7) leaves for special appointments; (8) resignations; (9) miscellaneous. Give individual's name, present place of employment, new title, new institution, and date on which change is to occur.

Deadlines for each issue are: Spring, December 15; Summer, March 15; Fall, June 15; Winter, September 15.

Please send all information to the *Journal of Economic Perspectives*, Department of Economics, Stanford University, Stanford, CA 94305-6072. We reserve the right to edit material received.

Accession Number.
121814
Date 7·12·89

The Journal of

Economic Perspectives

A journal of the American Economic Association

Editor
Joseph E. Stiglitz, Princeton University

Co-editor
Carl Shapiro, Princeton University

Associate Editors
Henry J. Aaron, The Brookings Institution
Stanley Fischer, Massachusetts Institute of Technology
Dwight M. Jaffee, Princeton University
Edward P. Lazear, University of Chicago and The Hoover Institution
Mark J. Machina, University of California at San Diego
Charles F. Manski, University of Wisconsin
John E. Roemer, University of California at Davis
Kenneth S. Rogoff, University of Wisconsin
Bernard Saffran, Swarthmore College
Steven C. Salop, Georgetown University Law Center
Lawrence H. Summers, Harvard University
Hal R. Varian, University of Michigan
Janet L. Yellen, University of California at Berkeley

Managing editor
Timothy Taylor

Editorial associate
Caroline Moseley

121814
7 12 89

Editorial offices:
Woodrow Wilson School of Public and International Affairs
Princeton University, Princeton, New Jersey 08544

The Journal of Economic Perspectives gratefully acknowledges the support of Princeton University.
Copyright © 1988 by the American Economic Association; All Rights Reserved.
Composed by Science Typographers, Inc., Medford, New York 11763, U.S.A.
Printed by Banta Company, Menasha, Wisconsin 54952, U.S.A.
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Winter 1988, Vol. 2, No. 1, is published quarterly (February, May, August, November) by the American Economic Association, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Annual subscription fees: Institutional subscriber–$125.00, Individual subscriber–$72.00, Regular members–$38.50, $46.20, or $53.90 depending on income. A subscription also includes the *American Economic Review* and the *Journal of Economic Literature*. In countries other than the U.S.A., add $16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

AMERICAN ECONOMIC ASSOCIATION

1989 ANNUAL MEMBERSHIP RATES

Membership includes:

—a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

● Regular members with annual incomes of $30,000 or less	$42.00
● Regular members with annual incomes above $30,000 but no more than $40,000	$50.40
● Regular members with annual incomes above $40,000	$58.80
● Junior members (available to registered students for three years only). Student status must be certified by your major professor or school registrar	$21.00
● In Countries other than the U.S.A., Add $16.00 to cover postage.	
● Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association)	$8.40

Please enter my membership for the following period:

☐ Jan.-Dec. ☐ April-March ☐ July-June ☐ Oct.-Sept.

First Name and Initial ____ Last Name ____ Suffix ____

Address Line 1 ____

Address Line 2 ____

City ____

State or Country ____ Zip/Postal Code ____

MAJOR FIELDS (TWO ONLY)

LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*. "Classification System for Books."

☐ ☐

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____

____

Authorized Signature ____

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

The Journal of
Economic Perspectives

4

Contents

Volume 2 • Number 1 • Winter 1988

Symposium

Takeovers

Hal R. Varian, "Symposium on Takeovers" 3
Andrei Shleifer and Robert W. Vishny, "Value Maximization and the Acquisition Process" 7
Michael C. Jensen, "Takeovers: Their Causes and Consequences" 21
Gregg A. Jarrell, James A. Brickley, and Jeffry M. Netter, "The Market for Corporate Control: The Empirical Evidence Since 1980" 49
F. M. Scherer, "Corporate Takeovers: The Efficiency Arguments" 69

Articles

Ronald I. McKinnon, "Monetary and Exchange Rate Policies for International Financial Stability: A Proposal" 83
Rudiger Dornbusch, "Doubts about the McKinnon Standard" 105
John Williamson, "On McKinnon's Monetary Rule" 113
Dale J. Poirier, "Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics" 121
John Rust, "Comment on Poirier: The Subjective Perspective of a 'Spiritual Bayesian'" 145
Adrian Pagan, "Comment on Poirier: Dogma or Doubt" 153
John Geweke, "Comment on Poirier: Operational Bayesian Methods in Econometrics" 159
Dale J. Poirier, "The Subjectivist Response" 167
Robert J. Thornton and Jon T. Innes, "The Status of Master's Programs in Economics" 171

Features

Barry Nalebuff, "Puzzles: Penny Stocks, Discount Brokers, Better Bidding, and More" 179
Bernard Saffran, "Recommendations for Further Reading" 187
Richard H. Thaler, "Anomalies: The Winner's Curse" 191
Notes 203

American Economic Association

Summer Minority Program

at Temple University

June 2, 1989–July 28, 1989

A New Economics Program with Exciting Features

An Attractive Stipend–plus Books, Room, Board, Tuition and Travel are Available to Qualified Students.

For More Information Contact:

DEPARTMENT OF ECONOMICS
RITTER HALL ANNEX
TEMPLE UNIVERSITY
PHILADELPHIA, PA 19122

(215-787-8887)

Announcement

Beginning with the next issue (Spring 1988) the *Journal of Economic Perspectives* will include a "Correspondence" feature based on letters from readers concerning the journal. Letters should be relatively short—generally less than 1000 words—and should be sent to the journal offices at the address appearing inside the front cover. The editors will select letters for publication, and all published letters will be subject to editing for style and length.

Announcement

The *Journal of Economic Perspectives* intends to publish an occasional review of recent papers studying the economics profession. Such papers may be recently published, soon-to-be published, or unpublished. The review would cover such topics as rankings of economics departments for whatever reason and by whatever criteria, studies of the market for hiring economists, trends in publication records, and so on. If you would like your research results to be considered for inclusion in the review, please send a copy of the full paper by April 1 to Professor David C. Colander, Department of Economics, Middlebury College, Middlebury, Vermont 05753.

Journal of Economic Perspectives
Advisory Board

Marcus Alexis, University of Illinois at Chicago
Elizabeth E. Bailey, Carnegie Mellon University
Gary S. Becker, University of Chicago
James M. Buchanan, George Mason University
David C. Colander, Middlebury College
Robert Eisner, Northwestern University
Jacob A. Frenkel, International Monetary Fund
Darius W. Gaskins, Burlington Northern Railroad
W. Lee Hansen, University of Wisconsin
Robert L. Heilbroner, New School for Social Research
Alice Rivlin, The Brookings Institution
Charles L. Schultze, The Brookings Institution
Robert M. Solow, Massachusetts Institute of Technology
James Tobin, Yale University

The Journal of
Economic Perspectives

Contents

Volume 2 • Number 2 • Spring 1988

Lecture

Martin S. Feldstein, "Distinguished Lecture on Economics in Government: Thinking about International Economic Coordination" 3

Symposia

Explaining Savings

Franco Modigliani, "The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth" 15
Laurence J. Kotlikoff, "Intergenerational Transfers and Savings" 41

Public and Private Unionization

Edward P. Lazear, "Symposium on Public and Private Unionization" 59
Richard B. Freeman, "Contraction and Expansion: The Divergence of Private Sector and Public Sector Unionism in the United States" 63
Melvin W. Reder, "The Rise and Fall of Unions: The Public Sector and the Private" 89

Articles

David M. Kreps, "In Honor of Sandy Grossman, Winner of the John Bates Clark Medal" 111
Richard Carson and Peter Navarro, "A Seller's (and Buyer's) Guide to the Job Market for Beginning Academic Economists" 137

Features

Barry Nalebuff, "Puzzles: Cider in Your Ear, Continuing Dilemma, The Last Shall be First, and More" 149
Bernard Saffran, "Recommendations for Further Reading" 157
Richard H. Thaler and William T. Ziemba, "Anomalies: Parimutuel Betting Markets: Racetracks and Lotteries" 161
Correspondence 175
Notes 185

Journal of Economic Perspectives

Advisory Board

Marcus Alexis, University of Illinois at Chicago
Elizabeth E. Bailey, Carnegie Mellon University
Gary S. Becker, University of Chicago
James M. Buchanan, George Mason University
David C. Colander, Middlebury College
Robert Eisner, Northwestern University
Jacob A. Frenkel, International Monetary Fund
Darius W. Gaskins, Burlington Northern Railroad
W. Lee Hansen, University of Wisconsin
Robert L. Heilbroner, New School for Social Research
Alice Rivlin, The Brookings Institution
Charles L. Schultze, The Brookings Institution
Robert M. Solow, Massachusetts Institute of Technology
James Tobin, Yale University

The Journal of

Economic Perspectives

Contents

Volume 2 • Number 3 • Summer 1988

Symposia

Brady Commission Report on the October 1987 Stock Market Crash

Bruce Greenwald and Jeremy Stein, "The Task Force Report: The Reasoning Behind the Recommendations" 3

James F. Gammill, Jr. and Terry A. Marsh, "Trading Activity and Price Behavior in the Stock and Stock Index Futures Markets in October 1987" 25

Hayne Leland and Mark Rubinstein, "Comments on the Market Crash: Six Months After" 45

Monetary Lessons of the 1980s

Benjamin M. Friedman, "Lessons on Monetary Policy from the 1980s" 51

William Poole, "Monetary Policy Lessons of Recent Inflation and Disinflation" . . . 73

Immigration

Barry R. Chiswick, "Illegal Immigration and Immigration Control" 101

Clark W. Reynolds and Robert K. McCleery, "The Political Economy of Immigration Law: Impact of Simpson-Rodino on the United States and Mexico" 117

Articles

Charles Brown, "Minimum Wage Laws: Are They Overrated?" 133

James H. Stock and Mark W. Watson, "Variable Trends in Economic Time Series" 147

Features

Joseph Farrell, "Puzzles: Sylvia, Ice Cream and more" 175

Bernard Saffran, "Recommendations for Further Reading" 183

Robyn M. Dawes and Richard H. Thaler, "Anomalies: Cooperation" 187

Correspondence 199

Notes 204

Announcement

The American Economic Association has established an ad hoc committee on the *Journal of Economic Perspectives*. William Baumol of Princeton University and New York University heads the committee; members are Leonid Hurwicz of the University of Minnesota, Michael McPherson of Williams College, Alfred Rees of the Alfred P. Sloan Foundation, and Robert Solow of the Massachusetts Institute of Technology. The task of the committee has two parts: (1) to recommend the financial terms on which the *Journal of Economic Perspectives* and/or the two established journals, the *American Economic Review* and the *Journal of Economic Literature*, should henceforth be provided to the membership, and (2) to comment on the contents and quality of the *Journal of Economic Perspectives*, paying particular attention to its correspondence with the desires of membership.

The committee does intend to carry out a small sample survey of AEA members. However, anyone with additional comments or ideas for the committee should send them as soon as possible to Professor William Baumol, New York University, Department of Economics, 269 Mercer Street, 7th Floor, New York, New York 10003. The committee plans to finish its report in time for the December 1988 AEA annual meetings.

Announcement

The *Journal of Economic Perspectives* has relocated its editorial offices. Its new address is: *Journal of Economic Perspectives*, Department of Economics, Stanford University, Stanford, CA, 94305-6072.

Journal of Economic Perspectives

Advisory Board

Marcus Alexis, University of Illinois at Chicago
Elizabeth E. Bailey, Carnegie Mellon University
Gary S. Becker, University of Chicago
James M. Buchanan, George Mason University
David C. Colander, Middlebury College
Robert Eisner, Northwestern University
Jacob A. Frenkel, International Monetary Fund
Darius W. Gaskins, Burlington Northern Railroad
W. Lee Hansen, University of Wisconsin
Robert L. Heilbroner, New School for Social Research
Alice Rivlin, The Brookings Institution
Charles L. Schultze, The Brookings Institution
Robert M. Solow, Massachusetts Institute of Technology
James Tobin, Yale University

27 DEC 1988

The Journal of

Economic Perspectives

Contents

Volume 2 • Number 4 • Fall 1988

Symposium

The Slowdown in Productivity Growth

Stanley Fischer, "Symposium on the Slowdown in Productivity Growth" 3

Zvi Griliches, "Productivity Puzzles and R&D: Another Nonexplanation" 9

Dale W. Jorgenson, "Productivity and Postwar U.S. Economic Growth" 23

Mancur Olson, "The Productivity Slowdown, the Oil Shocks, and the Real Cycle" 43

Michael J. Boskin, "Tax Policy and Economic Growth: Lessons From the 1980s" . . 71

Articles

Merton H. Miller, "The Modigliani-Miller Propositions After Thirty Years" 99

Joseph E. Stiglitz, "Why Financial Structure Matters" 121

Stephen A. Ross, "Comment on the Modigliani-Miller Propositions" 127

Sudipto Bhattacharya, "Corporate Finance and the Legacy of Miller and Modigliani" 135

Franco Modigliani, "MM—Past, Present, and Future" 149

Theodore Morgan, "Theory versus Empiricism in Academic Economics: Update and Comparisons" 159

Thomas Romer, "Nobel Laureate: On James Buchanan's Contributions to Public Economics" 165

Features

Barry Nalebuff, "Puzzles: Blockades, Carrier Missions, Secret Intelligence, and More" 181

Bernard Saffran, "Recommendations for Further Reading" 191

Richard H. Thaler, "Anomalies: The Ultimatum Game" 195

Correspondence 207

Notes 213

Journal of Economic Perspectives

Advisory Board

Marcus Alexis, University of Illinois at Chicago
Elizabeth E. Bailey, Carnegie Mellon University
Gary S. Becker, University of Chicago
James M. Buchanan, George Mason University
David C. Colander, Middlebury College
Robert Eisner, Northwestern University
Jacob A. Frenkel, International Monetary Fund
Darius W. Gaskins, Burlington Northern Railroad
W. Lee Hansen, University of Wisconsin
Robert L. Heilbroner, New School for Social Research
Alice Rivlin, The Brookings Institution
Charles L. Schultze, The Brookings Institution
Robert M. Solow, Massachusetts Institute of Technology
James Tobin, Yale University

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 3–5

Symposium on Takeovers

Hal R. Varian

The 1980s have been a decade of innovation in financial markets: there have been new financial products, a new degree of participation, a new increase in stock market value, and a new wave of mergers, acquisitions, and takeovers. The latter topic is the concern of this symposium.

Of the hundred largest mergers and acquisitions on record up to 1984, 65 occurred between 1981 and 1983, and only 11 occurred prior to 1979. During the 1981–84 period, there were at least 45 transactions of over a billion dollars apiece; prior to this period there were only a dozen or so transactions of such magnitude.

What is the meaning of this "merger mania?" Why did it occur so suddenly? Is the wave of takeovers merely a speculative bubble, or is it motivated by fundamental economic conditions? What can we expect in the future? What policy responses, if any, are appropriate in light of this explosion of takeover activity?

The papers in this symposium offer four perspectives on takeover activities. It is generally agreed that such activities are primarily a manifestation of the "market for corporate control." That is, takeovers serve as an incentive device for management to perform in the stockholders' interests. If the current managers fail to maximize value, a takeover offers a mechanism whereby shareholders can replace the current managers with managers who offer better performance. To the extent that acquisitions encourage efficient operations by management, they can be an effective mechanism for

■ *Hal R. Varian is Reuben Kempf Professor of Economics and Professor of Finance, University of Michigan, Ann Arbor, Michigan, where he is associated with the J. Ira Harris Center for the Study of Corporate Finance.*

enhancing economic performance. On the other hand, there are costs as well as benefits to takeovers, and a sound policy appraisal must take account of those costs as well.

Andre Shleifer and Robert Vishny begin by reviewing the received wisdom on the "separation of ownership and control," and the implications of this view on the performance of corporate management. After reviewing the evidence on internal control devices, they argue that "hostile takeovers are probably the most effective way for shareholders to get rid of non-value-maximizing managers." However, Shleifer and Vishny also point out that there are many problems with using takeovers as an incentive mechanism: a free-rider problem, strategic defense by management, non-value-maximizing motivations by the *acquiring* firms, the winner's curse, interference with long-term contracts with workers, and more. They conclude by offering several suggestions for change that they feel will make the market for corporate control operate more efficiently.

Michael Jensen provides a succinct review of the takeover controversy and the evidence to date. He argues, "The market for corporate control is creating large benefits for shareholders and for the economy as a whole by loosening control over vast amounts of resources and enabling them to move to their highest-value use." Jensen examines the arguments of many of the critics of takeovers and mergers and presents theoretical and empirical evidence opposing these views.

Jensen goes on to describe the role of "free cash flow" in explaining various aspects of takeovers. Free cash flow is defined as cash flow in excess of that required to fund all of a firm's projects that have positive net present values when discounted at the relevant cost of capital. Managers acting in the best interest of shareholders should distribute this cash flow to the shareholders rather than waste it through organizational inefficiencies or ill-advised investments. However, in some organizations managers may need the threat of a hostile takeover to provide the inducement to act in the shareholders' interests. Jensen describes a wide range of phenomena that the free cash flow theory can explain, including the industrial pattern of takeovers, the performance of acquired firms, the determinants of takeovers, and others. Finally, Jensen examines some of the mechanisms of takeovers, such as junk bonds, golden parachutes, greenmail, and poison pills, and appraises their contribution to economic efficiency.

Gregg A. Jarrell, James A. Brickley, and Jeffry M. Netter provide a comprehensive review of the empirical evidence regarding takovers since 1980. (The evidence before 1980 is described in detail in the 1983 Symposium on the Market for Corporate Control, published in the *Journal of Financial Economics*, *11*, nos. 1–4.)

They first examine the magnitude and the distribution of shareholder gains to acquisitions, and then turn to a discussion of the sources of such gains. Myopia, winners' curses, tax effects, redistribution from bondholders and workers, are all examined and, ultimately, rejected as sources of gains. The authors conclude that "the evidence is consistent with the notion that these corporate transactions reflect economically beneficial reshufflings of productive assets."

Jarrell, Brickley, and Netter then discuss the empirical evidence concerning various defenses against takeovers. They argue that such tactics typically hurt the shareholders of target firms and benefit the managers of such firms. Such findings may have an important role to play in the legal controversy surrounding hostile takeovers.

F. M. Scherer provides an alternative view of takeover activities. The other authors in this symposium generally make the assumption that the stock market valuation accurately reflects the true value of both the acquired and the acquiring firms. Scherer argues that this relatively strong hypothesis of market efficiency is at least open to doubt and suggests an alternative research strategy. In several previous papers, Scherer and David Ravenscraft have used the Line of Business data collected by the Federal Trade Commission to examine nearly 6,000 acquisitions between 1950 and 1976. Scherer found that on average the targeted firms were underperformers relative to other firms in their home industry, but that takeovers did not significantly improve economic performance. Using these data, "the hypothesis that takeovers *improve* performance is not supported."

Since the merger wave of the 1980s may differ from the earlier experience described in this study, Scherer discusses some more recent phenomena, including "bust-up takeovers," debt financing, and managerial decision-making horizons. He ends on a cautionary note: "Recent takeovers have probably accelerated desired divisional restructuring and led to lower-cost financial structures, but not without risks to the macroeconomy."

What are readers to make of this discussion of takeovers? I think that several conclusions emerge. First, it seems clear that the stock market responds positively to acquisitions—Wall Street, at least, thinks that shareholders gain from acquisitions. Second, it is difficult to identify with any precision the source of these gains and it is even possible, given the findings of Scherer and Ravenscraft, that some of the gains may be illusory. Third, even though the source of the gain is unclear, it seems clear that many of the defensive strategies of management may operate against the social interest. To the extent that more trade is better, social policies that restrict voluntary exchanges of assets are bad policies. This maxim applies just as well to financial markets as to other sorts of markets.

■ *The author would like to thank Michael Bradley for consultation on this project.*

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 7–20

Value Maximization and the Acquisition Process

Andrei Shleifer and Robert W. Vishny

Like the rest of us, corporate managers have many personal goals and ambitions, only one of which is to get rich. The way they try to run their companies reflects these personal goals. Shareholders, in contrast, deprived of the pleasures of running the company, only care about getting rich from the stock they own. Hence, when managers ignore profits to keep up traditional lines of business, conflicts are bound to arise. While many academic papers teach us that shareholder and market pressure will still force managers to maximize value,[1] the newspapers remind us that this is not always the case. Much corporate behavior seems best understood in terms of managers running the show largely as they please.

The takeover wave of the 1980s put the manager-shareholder conflict to a new test. Where other checks on management failed, hostile takeovers could now wrest control from managers who ignored the interests of their shareholders. More so than ever before, fear of such disciplinary takeovers has forced managers to listen to shareholder wishes. But even now, many acquisitions are not of this disciplinary variety. Ironically, making acquisitions is often just the quickest and easiest way for managers to expand the scope of their control by directing the firm's cash flows into new ventures.

In this paper, we appraise the acquisition process from the managerial perspective. Has the pressure brought by hostile takeovers effectively restricted non-

---

[1] By value maximization we mean maximization of the net present value of future profits.

■ *Andrei Shleifer and Robert W. Vishny are Assistant Professors of Finance, the Graduate School of Business, University of Chicago, Chicago, Illinois.*

value-maximizing conduct by managers? Are acquisitions themselves driven by non-value-maximizing behavior on the part of acquiring managers? We conclude with some recommendations for improving the takeover process.

Why Are Takeovers Needed? The Failure of Internal Controls

If a manager enjoys non-value-maximizing behavior enough, the first-best complete information employment contract would compensate him in the form of such behavior, rather than paying him to avoid it completely. With imperfect monitoring, the optimal contract would also leave a risk-averse manager some room for non-value-maximizing behavior, even if none of such behavior would be allowed in the first-best world. Incomplete information or a strong preference for non-value-maximizing behavior can thus explain the existence of such behavior even with optimal employment contracts.

But in practice, non-value-maximizing behavior seems vastly to exceed what one would expect to observe under optimal incomplete information contracts. A wealth of evidence suggests that managers often take actions that dramatically reduce the value of the firm. Witness, for example, the ability of managers to defeat hostile tender offers, even when shareholders stand to gain hundreds of millions of dollars from the bid, or the ability to undertake billion dollar oil exploration projects that the market treats as having negative present value (McConnell and Muscarella, 1985). Understanding the limitations of the internal control mechanisms of the firm will help us explain how hostile takeovers serve to reassert the interests of shareholders.

The board of directors is the main internal mechanism through which shareholders can try to constrain managerial choices. The board has at least nominal power to hire and fire the chief executive officer and to block any major corporate projects. In practice, however, boards of directors are rarely effective in stopping the non-value-maximizing behavior of managers. As a rule, managers control the selection of directors (Mace, 1971). The elected directors are often either insiders loyal to the firm's managers, or they are outsiders with a financial interest in the continuity of management, such as lawyers or advertisers for the firm. Even the relatively independent directors are often co-opted by the chief executive officer into putting the interests of managers and employees before those of shareholders. At the very least, directors are usually predisposed to give top managers the benefit of the doubt, rather than eager to find fault with their work. Although Warner, Watts and Wruck (forthcoming) and Weisbach (forthcoming) have found that poor stock performance is followed by higher CEO turnover—thus implying some monitoring by the board—the magnitude of the effect points to only a small employment risk for a top manager bent on non-value-maximizing behavior.

Even when the board of directors is determined to fulfill its fiduciary duty to shareholders and to work toward value maximization, it usually lacks sufficient knowledge to decide that a project is not value-maximizing, and thus has no basis on which to challenge the choices of managers. Collecting the necessary information can be very costly and it will rarely pay a director without a large ownership stake to do so. Nor are the managers themselves likely to provide the board with the knowledge required for effective monitoring. Among the constraints it faces, limitations on information are probably the most important obstacle to the board's governance.

In discussing the board of directors as a control device, it is important to recognize that a manager is likely to take whatever actions he can to make himself more valuable to the firm, and more costly to remove. Keeping valuable information from the board and from fellow executives is an example of this kind of activity. Perhaps a better way for the chief executive officer to promote self-entrenchment is by investing in the businesses he is good at running. A CEO with a railroad background will upgrade the railroad, rather than raise dividends, even if the railroad investment is not value-maximizing. After the investment is made, this manager might well be the best choice for the job and certainly not worth firing. If the board fails to prevent such investments, the top manager is secure at the helm. The board is at a particular disadvantage in criticizing major decisions about which businesses to enter and exit, since that is where detailed knowledge of a variety of businesses is essential.

A board might best be able to exercise its power at the point of managerial succession, when the future chief executive officer is not yet entrenched. For example, the board can pick a successor whose talents and preferences incline him to maximizing value, or it can write a contract that limits the new manager's options in increasing firm size or investing. However, choosing the next manager is widely viewed as the prerogative of the departing chief executive officer (Vancil, 1987). If the current top executive picks a replacement with an eye to continuing his own policies, he would not necessarily want to choose a value maximizer or to limit this successor's ability to reinvest cash flows.

Absent the possibility of dictating a course of action to the manager, the board can still design a compensation contract to provide incentives for value maximizing. Such a contract would take the form of a direct ownership stake in the firm (or more generally, stock options). Unfortunately, the legal environment in which the corporation operates usually limits large scale payments to executives. For example, take the situation of a chief executive officer in a company with a weak board of directors. This CEO can either engage in non-value-maximizing behavior, or else convince the board to compensate him with a large ownership position in the firm that aligns his interests with those of shareholders. If the CEO pursues the former strategy, he is protected from outside interference both by virtue of his ineffectual board, and by the "business judgment" rule that keeps the court from meddling in the affairs of the company.[2] If, in contrast, the CEO arranges for a valuable compensation contract, he can be sued by shareholders for breach of fiduciary duty and the business judgment rule no longer protects him. Since under the business judgment rule shareholders can only challenge clear conflicts of interest in court, actual compensation is naturally biased against stock or money and toward non-value-maximizing behavior. Such behavior then becomes a substitute, yet inefficient form of compensation.

---

[2] Gilson (1986) describes the business judgment rule as follows: absent bad faith or some other corrupt motive, directors are normally not liable to the corporation for mistakes of judgment, whether those mistakes are classified as mistakes of fact or mistakes of law. Gilson also quotes Manne (1967) who writes: "The liability aspects of the rule may well have been incidental to its principal function. The rule is more likely to have survived because it functioned as a quasi-jurisdictional barrier to prevent courts from exercising regulatory powers over the activities of corporate managers."

Even when the threat of shareholder lawsuits does not constrain the compensation of top officers, other considerations often do. For example, many outside directors think it unethical to "bribe" the top executive to do what he is supposed to do anyway, even when higher compensation is the only effective means of getting the manager to act for the benefit of the shareholders. Indeed, there is an economic justification for maintaining a rule of refusing to respond to a conflict of interest with a bribe. If executives can expect to receive generous payments whenever it looks like they may not act in the interest of shareholders, they will have an incentive to create such situations. Finally, many corporate boards feel constrained to set compensation of top executives by reference to an industry- or economy-wide standard, rather than experimenting with more innovative ways of compensating top officers.

As long as the chief executive officer does not place too much weight on non-value-maximizing behavior, even moderate-sized compensation contracts that will be approved by corporate boards and upheld in court can go some of the way toward preventing such behavior. If managers are rewarded for good stock performance, and penalized for poor stock performance, they will have an incentive to maximize value. Murphy (1985) finds that managers' pay responds positively to stock performance, while studies done by Brickley, Bhagat, and Lease (1985) and Tehranian and Waegelian (1985) show that stock prices rise on announcements of introduction of incentive-based compensation schemes. To interpret these studies as showing that managers are properly motivated to maximize value is, however, to overstate the case.

No study we know of indicates that the incentives provided by extant compensation contracts effectively discourage most non-value-maximizing behavior. For example, the increase in the price of shares observed in the studies of announcements of incentive schemes is much larger when the incentive contract is short term rather than long term. This finding suggests that managers get short-term contracts when they know that good earnings are coming, and the market recognizes this pattern by updating its earnings forecast on the announcement of the plan. The increased stock price reflects this update, not improved incentives. In fact, short-term incentive contracts are followed by announcements of unusually high earnings. Consistent with our scepticism about the efficacy of compensation arrangements, Jensen and Murphy (1986) find that the lifetime wealth of the average chief executive officer of a large firm increases by only about $1.40 for each $1000 increase in the market value of the firm.

In sum, internal control devices are not especially effective in forcing managers to abstain from non-value-maximizing conduct. On the other hand, persuading better-informed managers to limit non-value-maximizing conduct might be very expensive, and possibly infeasible if considerations of fairness and moral hazard preclude the necessary compensation arrangements. In these circumstances, it is not surprising that external means of coercion such as hostile takeovers can come to play a role.

Hostile Takeovers as Disciplining Devices

In light of the lackluster performance of other disciplining devices, hostile takeovers are probably the most effective way for shareholders to get rid of non-value-maximizing managers without bribing them. In these transactions, a takeover specialist (raider) or a second company bids directly for the shares of the target firm. By doing so, the bidder deals directly with the target's shareholders, rather than with its management. A bidder who gains the requisite votes then assumes control and removes the incumbent managers. He then often increases corporate debt, reduces employment, tightens management contracts and sells off parts of the company. The resulting increases in profitability enable acquirers to pay 30 percent to 50 percent premiums in these deals.

Hostile acquisitions have the elements of both coercion and persuasion. On the one hand, by going over the heads of incumbent target managers straight to shareholders, the bidder is able to remove these managers against their will. Where internal controls have failed, takeovers can often succeed. On the other hand, hostile takeovers are often accompanied by large separation payments to managers of target companies ("golden parachutes"), the purpose of which is to attenuate the opposition of managers to the bid. When top managers are losing their jobs, boards seem to find it more acceptable to give them large bribes. Even so, golden parachutes have often been challenged in court, if not always successfully.

Recent work by Morck, Shleifer, and Vishny (forthcoming) suggests that targets of hostile bids have many characteristics indicative of the need for external discipline. They consider a sample of 371 firms from the 1980 Fortune 500, of which 40 have been acquired between 1981 and 1985 in a contest that was at least initially hostile (many firms escaped into the hands of a friendly acquirer or were ultimately bought by a group of investors including incumbent managers). The average 1980 Tobin's q of a target of a future hostile bid is .524, as compared to the average q of .848 for the sample as a whole.[3] The ability of low q to predict which firm will be the target of a hostile takeover seems to come largely from the firm's being in a low q industry, rather than from just having an especially low q within its industry. Moreover, targets of hostile bids seem to be older firms that are investing a lower fraction of corporate income than does the average Fortune 500 firm.

These findings suggest that hostile takeovers affect industries in decline or sharp change where managers fail to shrink operations rapidly enough or to make other adjustments. In maintaining full-scale operations, managers may be guarding the domain of their control or trying to protect employees from dismissal or wage cuts. In the case of the oil industry, managers continued costly explorations for additional reserves despite declining oil prices (Jacobs, 1986; Jensen, 1986). In the case of airlines, managers were unable or unwilling to cut union wages that had reached very high levels under regulation.

---

[3] Tobin's "q" used in this study is the ratio of market value of the firm's debt and equity to the replacement cost of its physical capital.

The example of the declining or quickly changing industry suggests an interesting interpretation of the role of raiders. These takeover specialists acquire companies in a wide range of industries, and usually sell off some parts and curtail the operations of others. It is hard to believe that, say, Carl Icahn can have a simultaneous comparative advantage at running a textile mill, a railcar leasing company, a steel producer and an airline. It seems more plausible that raiders have a comparative advantage at transferring wealth from employees, suppliers and managers of the company to the shareholders. In other words, they squeeze other beneficiaries of corporate wealth in situations where the incumbent is unable or reluctant to do so.

Effectiveness of Hostile Takeovers

In their seminal paper, Grossman and Hart (1980) pointed to the free-rider problem in making a successful tender offer for a firm. They observed that small shareholders of the firm targeted for a takeover have the option of not tendering their shares and staying on as shareholders after management has changed. If they do not sell, they can share in the increase of the profitability of the firm resulting from policies of the new management. Since no single small shareholder can affect the outcome of the bid, the dominant strategy is not to tender unless the value of the bid fully reflects the expected increase in profitability under new management. But if the bidder has to surrender all potential gains to the current shareholders of the firm, he cannot gain from the acquisition. This reasoning implies that if acquiring information to make a bid or bidding itself is costly, no bids will take place.

In practice, this result says more about the difficulty of mounting a profitable bid than about the impossibility of doing so. Bids can in fact be profitable for a number of reasons. Grossman and Hart stress the ability of the bidder to divert some of the value gains his takeover produces. If the bidder need not share these gains with non-tendering shareholders, they will have an incentive to tender and the bidder can make money. Another way to mount a profitable bid, emphasized by Shleifer and Vishny (1986), is for the bidder to accumulate secretly some shares on the open market at a price that reflects the value of these shares under the old management. In this case, the bidder can still realize the full value of any managerial improvements on the shares so accumulated (5 to 10 percent under the current regulation and practice) and hence can profit from the bid. While these results show why making bids is not prohibitively costly, they also suggest that the bidders have to share a lot of the value gains with shareholders of the target firm. With high transaction costs, hostile takeovers can punish only the most dramatic cases of non-value-maximizing behavior.

Managers also have many ways of protecting themselves from hostile bids. Takeover defenses employed by target management range from the relatively innocuous fair price amendment to the practically insurmountable poison pill. The term "poison pill" refers to a class of defenses that impose significant costs on the bidder by diluting his equity holdings, revoking his voting rights or forcing him to assume unwanted financial obligations upon "triggering" the pill. The pill is typically triggered in the event of a tender offer or the accumulation of more than some specified percentage of the target's shares by a single holder. Unlike many other takeover defenses, the poison pill can be adopted without shareholder approval. The latter feature of the pill, along with its potential to exact significant punishment on potential acquirers, has made the pill perhaps the most popular takeover defense recently. Of the 380 poison pill adoptions in the exhaustive sample of Ryngaert (1987), over eighty percent took place in 1986.

We are not suggesting that the hostile takeovers of the 1980s had no effect on non-value-maximizing behavior of managers. Almost 10 percent of the 1980 Fortune 500 have since been acquired in a transaction that started as hostile (and ended either as a hostile acquisition, an escape to a friendly white knight, or a management buyout). Managers of many other companies were probably scared enough to restrain their deviations from value maximization. In fact, some have beaten potential acquirers to the punch by taking on more debt and selling off divisions to escape a hostile bid or to deter one.

While some managers responded to the threat of a hostile bid by implementing the very changes sought by raiders, other managers only dug in deeper. For example, many managers changed their corporate charters in a way that makes them virtually invulnerable to any control challenge, whether through a takeover or through the proxy mechanism. Many state laws have also been rewritten to favor incumbent management. While non-value-maximizing behavior declined in some companies, in others such behavior became even harder to penalize.

Do Acquisitions Serve Managers or Shareholders?

Ironically, the literature that focused on takeovers as devices to eliminate non-value-maximizing behavior has almost completely forgotten the bidders, despite the fact that acquisitions may be the most important decisions about the allocation of corporate wealth that managers make. Acquisitions, especially friendly ones, may provide managers their greatest opportunity for expressing their non-value-maximizing preferences.

The acquisition process is probably the most important vehicle by which managers enter new lines of business. In his sample of 33 large diversified U.S. corporations, Porter (1987) finds that between 1950 and 1986 his firms entered an average of 80 new industries each and that over 70 percent of this diversification was accomplished through acquisition. Large scale movement of U.S. manufacturing toward unrelated diversification is now thought by many observers (including Porter) to have been unsuccessful. The high level of divestiture of peripheral businesses by diversified corporations beginning in the mid-1970s is almost surely a response to that failure.

As mentioned before, the decisions about which lines of business to enter and which companies to pursue are to a large extent the prerogative of the chief executive officer. In making these decisions, managers often feel that they have something to add to the firms they buy, be it a technological synergy, access to new markets, access to capital, or their own talents (Donaldson and Lorsch, 1983). But value maximization is not the only objective of the manager. The CEO also has ideas about the size and the character of the organization he wants to lead. For example, if the company operates in a declining industry, the CEO might want to start moving into faster growing industries. Many corporate acquisitions seem to be governed by this desire of managers to switch into businesses with long term growth potential even when the managers have no special expertise in running such businesses and when the value-maximizing strategy is to distribute free cash flows to shareholders.

Since in choosing the acquisition targets the manager is guided by a number of objectives other than value maximization, he is likely to overpay relative to what the acquisition is worth to shareholders. For in addition to the value gained for shareholders, the manager is also paying for increasing the size of the firm, the opportunity to diversify, and for making himself less replaceable. Examples of acquisitions that were probably motivated by non-value considerations include U.S. Steel's white knight acquisition of Marathon Oil, Mobil's acquisition of Montgomery Ward, Exxon's acquisition of Reliance Electric, and GM's acquisition of Hughes Aircraft. Of course, some managers may raise profitability of companies they acquire even when value-maximization is not their main motivation. Even managers interested only in entrenching themselves will enter businesses to which they have something to add, since this would increase their own value to shareholders.

On these points we differ with Roll's (1986) theory, in which managers infected by hubris try to maximize value, but overestimate the value of what they buy. As a result, they simply overpay. While we agree with Roll's belief that managers overpay, our reasoning is different. It is not that managers make valuation mistakes, but rather that they pay for the benefits of the acquisition that they care about but shareholders do not.

The implication of our story is that friendly mergers can (though they need not) create value, but that in many cases more than 100 percent of these gains will accrue to target shareholders. Shareholders of the target company gain and those of the acquiring company lose, since the latter end up paying the former for the benefits going to acquiring managers. But more than just a transfer between acquiring and target shareholders is likely to be involved here. Acquisition choices based on managerial objectives will not lead to the optimal allocation of managers to businesses. For example, managers of cash cows in declining industries will end up buying businesses that could be run more efficiently by other potential acquirers.

The evidence on the returns to shareholders of bidding firms and on changes in the combined market value of the two firms in a merger is unfortunately ambiguous. Jensen and Ruback (1983) interpret the studies they survey as showing that shareholders of the bidding firms in mergers do not lose. Roll (1986) cites results showing that they do lose on average on the announcement of the merger, but that such losses are often not statistically significant. Bidding firms seem to have done better recently, as the popularity of conglomerate mergers declined and the leniency of the Justice Department toward potentially anticompetitive mergers increased. Because bidding firms are much larger than the takeover targets in most cases, large positive target returns in conjunction with near zero bidder returns do not necessarily imply a significant increase in the combined value of the merging firms. Still, the combined gains from takeovers appear to be positive, although Roll notes that this conclusion is often not statistically reliable.

In our interpretation of the acquisition process, non-value-maximizing behavior of bidders plays a central role. The willingness of the managers of the bidder to pay for benefits to themselves that are of no value to their shareholders explains negative returns to acquiring firms. This interpretation suggests that before stressing the role of takeovers in eliminating non-value-maximizing behavior by managers of target companies, it is important to remember the managers of bidding firms. For them, the purchase of other companies at inflated prices may be the grandest deviation from value maximization. Just as improvements in acquisition techniques of the 1980s have pressured some managers to maximize value, they have enabled others to deviate from doing so on a previously impossible scale.

The Efficiency of Takeovers

The variety of takeover experiences makes an overall welfare analysis complicated. In many cases, it is difficult to identify the real or even perceived sources of efficiency gains that are to result from a takeover. Moreover, besides changing the size of the pie, most acquisitions also alter its division. Most transactions have both gainers and losers.

Much of non-value-maximizing behavior of managers of hostile targets seems to transfer corporate wealth from shareholders to other nonmanagement constituencies, such as employees, suppliers and customers. Regime changes following takeovers redistribute wealth from these constituencies back to shareholders of the target and to the bidder. Shleifer and Summers (forthcoming) look at the sources of gains in Icahn's takeover of TWA. They show that the takeover premium paid to TWA shareholders is equal to about half the present value of wage losses to members of three TWA unions: pilots, flight attendants, and machinists.[4] While such transfers benefit shareholders and raiders, they cannot be counted as increases in the size of the social pie. Efficiency can improve, as would be the case if union wages were reduced to reflect value marginal product, but this efficiency gain is second order relative to the transfer.

Shleifer and Summers emphasize an important implication of the pervasiveness of transfers in the takeover process. Some analyses have gauged the efficiency

[4] Of course, the takeover premium will at most reflect the *unexpected* component of the value of any wage cuts. In this case, the market may have been counting on current management or a raider reducing wages with high probability. In addition, one must take account of the fact that some portion of the gains from reducing wages probably went to Carl Icahn as a rent to his ability as a raider.

consequences of acquisitions by looking at changes in shareholder wealth. This is inappropriate. If acquisitions entail large transfers, analysts must look at wealth changes of all the relevant parties, including employees, suppliers and managers. For example, if some of shareholder gains in hostile takeovers come from reducing wages, then the calculation of social gains must take account of employee losses. The unfortunate implication of this analysis is that it becomes impossible to gauge the social consequences of acquisitions from shareholder returns alone.

Our emphasis on transfers in hostile takeovers may appear misplaced. If employees and managers have no legitimate property rights to the cash flows of the firm other than their explicitly stated compensation during the time of their contract, then transfers from them cannot be judged more harshly than transfers of stolen property from a thief. That is, if managers take for themselves and for other corporate constituencies what properly belongs to shareholders, then takeovers only correct this wrong. This claim that shareholders have the property right to all the firm's cash flows not allocated by an explicit contract is at the heart of many endorsements of hostile takeovers.

The trouble with this argument is that shareholders do not buy an asset that entitles them to the maximum feasible profits. Rather, they buy an asset whose price reflects the weaknesses of internal and external control mechanisms and the consequent non-value-maximizing behavior of managers. While shareholders of some firms get a raw deal when managers direct an unexpectedly large share of the cash flows to meet their personal goals, in other firms shareholders get more than they counted on when control devices operate better than expected. Because managerial actions are not fully specified in the contract, shareholders are in effect buying a (fairly priced) lottery ticket over the control related outcomes that determine production and dividends. This ticket does not entitle them to a takeover with probability one whenever management does not deliver the maximum feasible profit.

In fact, takeovers that radically change the environment from what shareholders, managers and employees expect may be objectionable, since they can harm parties acting on the basis of an implicit contract. For example, if managers or employees counted on a guarantee of employment with the company and planned their consumption and savings accordingly, then violating their expectations is a cost. Respecting expectations is to some extent an end in itself. One way to counter these unexpected transfers is to require shareholders to compensate, at least partially, displaced managers and employees. The government could also help out by providing benefits to the losers in hostile takeovers, especially if one thinks that such transactions should be subsidized as they create large efficiency gains that cannot be appropriated by the raider.

Increased availability of junk bond financing, the leniency of the Reagan administration antitrust stance, and the deregulation of transportation and banking have all contributed to a surprise change in the takeover environment. The consequences of such an unexpected change are easiest to justify when the gains to shareholders net of employee and management losses are large relative to the harm done from invalidating the implicit contract between them. The compensation principle from welfare economics can then be used to justify the transaction. Such gains might be especially large when side payments to managers are infeasible, so that shareholder gains from eliminating non-value-maximizing behavior far exceed the private benefits to management from engaging in such behavior. The gains might also be large when an employee layoff following a takeover substantially reduces the wage bill of the company, but most of the laid-off workers obtain employment at comparable wages elsewhere.

In evaluating takeovers from a longer-run perspective, shareholder gains net of losses (or gains) to other constituencies should be compared with the long-run costs of the process. One of the frequently mentioned costs is that of lawyers, investment bankers and other parties organizing the transaction. Jensen (1984) shows that these costs are small compared to the total market value gains (at least in hostile transactions). Furthermore, these costs presumably reflect in large part the superior ability of participating lawyers or bankers to capture rents. Since the social opportunity cost of the lawyers' and bankers' time is probably their social value in other rent-seeking activities, we think that the fees paid significantly overstate the true social cost of the process.

Other costs of takeover activity are potentially more serious. An increase in the incidence of takeovers makes it harder for managers at all levels to rely on long-term employment with the firm and to plan their activities accordingly. As a result, they will make fewer investments in their firm-specific human capital and undertake fewer long-term projects. They will also become less loyal to the firm, and less likely to work hard in hopes of future rewards. Some of the harm from decreased reliance on implicit contracts may be mitigated by switching to more explicit contractual arrangements for things like severance pay. But more explicit contracts will be only an imperfect substitute for managers who rely on the continuity of leadership and of the direction of corporate strategy to plan projects and to make firm-specific investments.

At the very highest levels of management, anxiety that the market's misunderstanding of the firm's strategy will result in a low stock price and a takeover can push investment choices toward those projects that yield the most immediately obvious payoffs, and away from those with the highest long-run payoffs. The focus on the short run at the expense of long-term investments can be particularly damaging to shareholders, as discussed by Stein (forthcoming).

Unfortunately, while shareholder benefits from hostile takeovers stare us in the face in the form of share price increases, the costs of takeovers are not so easily measured. For example, there is not enough evidence to reject the view that the lion's share of shareholder benefits from hostile takeovers comes from pure value transfers, such as union wage reductions with little change in production or employment. Yet our impression is that in at least some cases, hostile takeovers do force managers to move to a more productive (and not just more profitable) use of labor and capital. For example, in basic industries hurt by foreign competition such as steel and textiles, hostile takeovers have facilitated the needed wage concessions before financial distress disrupted otherwise viable operations. Takeovers have also forced reductions in reinvestment rates when even optimistic labor cost projections did not justify adding capacity to the industry. In the oil industry, for instance, hostile takeovers have prompted curtailment of excessive exploration. And bust-up takeovers of conglomerates have led to sales of peripheral divisions to managers who know them better.

But using takeovers to accelerate the process of adjustment in industries experiencing adverse shocks has its costs as well. In the face of an imminent takeover threat, managers respond with actions very different from those that are part of a long-run optimal retrenchment strategy. The very short planning horizon imposed by a takeover threat can cause panic, disruptions and waste. Even if, *ceteris paribus*, having takeovers is better than banning them, changing the corporate structure so as to avoid relying on takeovers is a still better way of dealing with manager-shareholder conflicts in declining industries. In our concluding section, we suggest some ways to improve the functioning of the internal controls of the firm. Ideally, this would result in lighting a fire under top managers without scorching the firm in the process.

Some Suggestions for Change

The analysis of this paper has stressed the importance of managerial preference for non-value-maximizing behavior and the limitations of the various mechanisms designed to bring manager and shareholder interests closer together. When internal control devices fail, external controls such as hostile takeovers can be effective, but they often entail disruptive changes in management. Whatever form controls take, they often lead to large unexpected transfers from other constituencies of the firm to shareholders. Social policy should then aim to improve the functioning of internal controls, and to alleviate the hardship caused by readjustment, especially in the case of takeovers.

First, the existing constraints on providing managers with share ownership or other compensation that might encourage them to serve shareholders should be removed. Courts should establish the precedent of blocking shareholder lawsuits against excessive compensation in cases where such compensation is necessary to improve the incentives of managers. Granting an ownership stake might be especially critical in the case of a declining firm, where divestment and contraction might be in shareholder interest but against the wishes of the managers concerned with the welfare of subordinates and with the size of the firm. In some cases, like that of Lee Iacocca, large ownership stakes have been given as an incentive device to new managers joining bankrupt firms. We hope that this practice will become acceptable even for firms that are not near bankruptcy. At the same time, courts should try to weaken the force of the business judgment rule as a shield against shareholder challenges to managerial decisions. While we recognize the virtues of the business judgment rule as a guard against disruption of corporate business, the current bias it fosters toward non-value-maximizing behavior is extreme.

Our second suggestion is that the board of directors should be compensated with stock. Upon being chosen for the board, each director should be given in lieu of his salary the amount of stock equal to what would now be five years worth of director compensation. Directors should be forbidden to sell this stock until they leave the board. While this stock holding might appear small, it is vastly larger than the current ownership of the majority of directors. Directors with relatively large ownership positions would be more likely to insist on value maximization and less likely to be co-opted by the chief executive officer. Taken together, these two suggestions should have the effect of reducing the non-value-maximizing behavior of managers, which in turn should lead to fewer tempting takeover targets and fewer acquisitive top managers running amok.

Third, if it is agreed that hostile takeovers accelerate the necessary adjustment in declining industries to a new economic environment—a judgment that still requires conclusive empirical confirmation—then we should have policies to smooth out such adjustments. In particular, severance pay and provisions to protect labor can alleviate the hardships of employees harmed by acquisitions and make implicit contracts more reliable. Shareholders of target firms should probably pay for at least some of these transfers to affected employees, but the government should also consider subsidizing this activity. Provisions to help employees would also deprive embattled managers of one of their favorite public excuses for opposing hostile bids. The way to deal with transfers that occur in hostile takeovers is to compensate the losers rather than to prohibit altogether acquisitions that might raise efficiency. Prohibiting hostile takeovers altogether would simply throw out the baby with the bathwater.

■ *The authors are grateful to Sidney Davidson, Carl Shapiro, Lawrence Summers and Timothy Taylor for helpful comments and advice.*

References

Brickley, James A., Sanjai Bhagat, and Ronald C. Lease, "The Impact of Long-Range Managerial Compensation Plans on Shareholder Wealth," *Journal of Accounting and Economics*, April 1985, *7*, 115–130.

Donaldson, Gordon, and Jay W. Lorsch, *Decision Making at the Top*. New York: Basic Books, 1983.

Gilson, Ronald J., *The Law and Finance of Corporate Acquisitions*. Mineola, NY: The Foundation Press, 1986.

Grossman, Sanford J., and Oliver D. Hart, "Takeover Bids, the Free-Rider Problem, and the Theory of the Corporation," *Bell Journal of Economics*, Spring 1980, *11*, 42–64.

Jacobs, E. Allen, "The Agency Costs of Corporate Control," MIT Working Paper, 1986.

Jensen, Michael C., "Agency Costs of Free Cash Flow, Corporate Finance and Takeovers," *American Economic Review*, Papers and Proceedings, May 1986, 323–329.

Jensen, Michael C., "The Takeover Controversy: Analysis and Evidence." In Coffee, John, Louis Lowenstein, and Susan Rose-Ackerman, eds., *Takeovers and Contests for Corporate Control*. New York. Oxford University Press, forthcoming.

Jensen, Michael C., "Takeovers: Folklore and Science," *Harvard Business Review*, November/December 1984, 109–121.

Jensen, Michael C., and Kevin J. Murphy, "Are Executive Compensation Contracts Structured Properly?" Managerial Economics Research Paper 86-14, Simon School of Management, 1986.

Jensen, Michael C., and Richard S. Ruback, "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics*, April 1983, *11*, 5–50.

Mace, Myles L., *Directors: Myth and Reality*. Boston: Harvard Business School, 1971.

Manne, Henry L., "Our Two Corporation Systems: Law and Economics," *Virginia Law Review*, 1967, *53*.

McConnell, John J., and Chris J. Muscarella, "Corporate Capital Expenditure Decisions and the Market Value of the Firm," *Journal of Financial Economics*, September 1986, *14*, 399–422.

Morck, Randall, Andrei Shleifer, and Robert W. Vishny, "Characteristics of Targets of Hostile and Friendly Takeovers." In Auerbach, Alan J., ed., *Takeovers: Causes and Consequences*. Chicago: University of Chicago Press, forthcoming.

Murphy, Kevin J., "Corporate Performance and Managerial Remuneration: An Empirical Analysis," *Journal of Accounting and Economics*, April 1985, 11–42.

Porter, Michael E., "From Competitive Advantage to Corporate Strategy," *Harvard Business Review*, May–June 1987, 43–59.

Roll, Richard, "The Hubris Hypothesis of Corporate Takeovers," *The Journal of Business*, April 1986, *59*, 197–216.

Ryngaert, Michael, "The Effect of Poison Pill Securities on Shareholder Wealth," Mimeo, University of Chicago, 1987.

Shleifer, Andrei, and Lawrence H. Summers, "Breach of Trust in Hostile Takeovers." In Auerbach, Alan J., ed., *Corporate Takeovers: Causes and Consequences*. Chicago: University of Chicago Press, forthcoming.

Shleifer, Andrei, and Robert W. Vishny, "Large Shareholders and Corporate Control," *Journal of Political Economy*, June 1986, *94*, 461–488.

Stein, Jeremy, "Takeover Threats and Managerial Myopia," *Journal of Political Economy*, forthcoming.

Tehranian, Hassan, and James F. Waegelian, "Market Reaction to Short Term Executive Compensation Plan Adoption," *Journal of Accounting and Economics*, April 1985, 7, 131–144.

Vancil, Richard F., *Passing the Baton*. Cambridge, MA: Harvard Business Press, 1987.

Warner, Jerold, Ross Watts, and Karen Wruck, "Stock Prices, Event Prediction, and Event Studies: An Examination of Top Management Changes," *Journal of Financial Economics*, forthcoming.

Weisbach, Michael S., "Outside Directors and CEO Turnover," *Journal of Financial Economics*, forthcoming.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 21–48

Takeovers: Their Causes and Consequences

Michael C. Jensen

Economic analysis and evidence indicate that the market for corporate control is benefiting shareholders, society, and the corporate form of organization. The value of transactions in this market ran at a record rate of about $180 billion per year in 1985 and 1986, 47 percent above the 1984 record of $122 billion. The gains to shareholders from these transactions have been huge. The gains to selling firm shareholders from mergers and acquisition activity in the ten-year period 1977–86 total $346 billion (in 1986 dollars).[1] The gains to buying firm shareholders are harder to estimate, and no one to my knowledge has done so as yet, but my guess is that they will add at least another $50 billion to the total. These gains, to put them in perspective, equal 51 percent of the total cash dividends (valued in 1986 dollars) paid to investors by the entire corporate sector in the past decade.[2]

These corporate control transactions and the restructurings that often accompany them are frequently wrenching events in the lives of those linked to the involved

[1] Estimated from data in Grimm (1986). Grimm provides total dollar values for all M & A deals for which there are publicly announced prices amounting to $500,000 or 10 percent of the firm in which at least one of the firms was a U.S. company. Grimm also counts in its numerical totals deals with no publicly announced prices that it believes satisfy this criteria. I assumed that the deals with no announced prices were on average equal to 20 percent of the size of the announced transactions.

[2] Total dividend payments by the corporate sector, unadjusted for inflation, are given in Weston (1986, p. 649). I extended these estimates to 1986.

■ *Michael C. Jensen is Professor of Business Administration, Harvard Business School, Cambridge, Massachusetts, and LaClare Professor of Finance and Business Administration and Director of the Managerial Economics Research Center, William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York.*

organizations: the managers, employees, suppliers, customers and residents of surrounding communities. Restructurings usually involve major organizational change (such as shifts in corporate strategy) to meet new competition or market conditions, increased use of debt, and a flurry of recontracting with managers, employees, suppliers and customers. This activity sometimes results in expansion of resources devoted to certain areas and at other times in contractions involving plant closings, layoffs of top-level and middle managers, staff and production workers, and reduced compensation.

Those threatened by the changes that restructuring brings about argue that corporate restructuring is damaging the American economy, damaging the morale and productivity of organizations, and pressuring executives to manage for the short-term. Further, they hold that the value restructuring creates does not come from increased efficiency and productivity; instead, the gains comes from lower tax payments, broken contracts with managers, employees and others, and mistakes in valuation by inefficient capital markets. Since the benefits are illusory and the costs are real, they argue, takeover activity should be restricted.

The controversy has been accompanied by strong pressure on regulators and legislatures to enact restrictions that would curb activity in the market for corporate control. Dozens of congressional bills in the last several years have proposed new restrictions on takeovers, but none have passed as of this writing. The Business Roundtable, composed of the chief executive officers of the 200 largest corporations in the country, has pushed hard for restrictive legislation. Within the past several years the legislatures of New York, New Jersey, Maryland, Pennsylvania, Connecticut, Illinois, Kentucky, Michigan, Ohio, Indiana and Minnesota have passed antitakeover laws. The Federal Reserve Board implemented new restrictions in early 1987 on the use of debt in certain takeovers.

In all the controversy over takeover activity, it is often forgotten that only 40 (an all-time record) out of the 3,300 takeover transactions in 1986 were hostile tender offers. There were 110 voluntary or negotiated tender offers (unopposed by management) and the remaining 3,100-plus deals were also voluntary transactions agreed to by management, although this simple classification is misleading since many of the voluntary transactions would not occur absent the threat of hostile takeover. A major reason for the current outcry is that in recent years mere size alone has disappeared as an effective takeover deterrent, and the managers of many of our largest and least efficient corporations now find their jobs threatened by disciplinary forces in the capital markets.

Economists have accumulated considerable evidence and knowledge on the effects of the takeover market. Most of the earlier work is well summarized elsewhere (Jensen and Ruback, 1983; Jensen, 1984; Jarrell, Brickley and Netter in this symposium). Here, I focus on current aspects of the controversy. In brief, the previous work tells us the following:

1. Takeovers benefit shareholders of target companies. Premiums in hostile offers historically exceed 30 percent on average, and in recent times have averaged about 50 percent.

2. Acquiring-firm shareholders on average earn about 4 percent in hostile takeovers and roughly zero in mergers, although these returns seem to have declined from past levels.
3. Takeovers do not waste credit or resources. Instead, they generate substantial gains: historically, 8 percent of the total value of both companies. Those value gains represent gains to economic efficiency, not redistribution between various parties.
4. Actions by managers that eliminate or prevent offers or mergers are most suspect as harmful to shareholders.
5. Golden parachutes for top-level managers do not, on average, harm shareholders.
6. The activities of takeover specialists (such as Icahn, Posner, Steinberg, and Pickens) benefit shareholders on average.
7. Merger and acquisition activity has not increased industrial concentration. Indeed, over 1,200 divestitures valued at $59.9 billion occurred in 1986, also a record level (Grimm, 1986).
8. Takeover gains do not come from the creation of monopoly power.

The market for corporate control is creating large benefits for shareholders and for the economy as a whole by loosening control over vast amounts of resources and enabling them to move more quickly to their highest-valued use. This is a healthy market in operation, on both the takeover side and the divestiture side, and it is playing an important role in helping the American economy adjust to major changes in competition and regulation of the past decade.

The Market for Corporate Control

The market for corporate control is best viewed as a major component of the managerial labor market. It is the arena in which alternative management teams compete for the rights to manage corporate resources (Jensen and Ruback, 1983). Understanding this point is crucial to understanding much of the rhetoric about the effects of hostile takeovers.

Managers often have trouble abandoning strategies they have spent years devising and implementing, even when those strategies no longer contribute to the organization's survival. Such changes can require abandonment of major projects, relocation of facilities, changes in managerial assignments, and closure or sale of facilities or divisions. Takeovers generally occur because changing technology or market conditions require a major restructuring of corporate assets, and it is easier for new top-level managers with a fresh view of the business and no ties with current employees or communities to make such changes. Moreover, normal organizational resistance to change is commonly significantly lower early in the reign of new top-level managers. For example, the premium Carl Icahn was able to offer for TWA and his victory over Texas Air for the acquisition of TWA were made possible in part by the willingness of TWA unions to negotiate favorable contract concessions with

Icahn—concessions that TWA management was unable to win prior to the takeover conflict. On the other hand, lack of detailed knowledge about the firm poses risks for new managers and increases the likelihood of mistakes.

A variety of political and economic conditions in the 1980s have created a climate where economic efficiency requires a major restructuring of corporate assets. These factors include the relaxation of restrictions on mergers imposed by the antitrust laws, withdrawal of resources from industries that are growing more slowly or that must shrink, deregulation in the financial services, oil and gas, transportation, and broadcasting markets, and improvements in takeover technology, including a larger supply of increasingly sophisticated legal and financial advisers, and improvements in financing technology such as the strip financing commonly used in leveraged buyouts and the original issuance of high-yield non-investment-grade bonds.

Each of these factors has contributed to the increase in total takeover and reorganization activity. Moreover, the first three factors (antitrust relaxation, exit, and deregulation) are generally consistent with data showing the intensity of takeover activity by industry. For example, the value of merger and acquisition transactions by industry in the period of 1981-84 given in Table 1 indicates that acquisition activity was highest in oil and gas, followed by banking and finance, insurance, food processing, and mining and minerals. For comparison purposes, the last column of the table presents data on industry size measured as a fraction of the total value of all firms. All but two of the industries, retail and transportation, represent a larger fraction of total takeover activity than their representation in the economy as a whole, indicating that the takeover market is concentrated in particular industries, not spread evenly throughout the corporate sector.

Many sectors of the U.S. economy have been experiencing slowing growth and, in some cases, even retrenchment. This phenomenon has many causes, including substantially increased foreign competition. This slow growth has increased takeover activity because takeovers play an important role in facilitating exit from an industry or activity. Major changes in energy markets, for example, have required a radical restructuring and retrenchment in that industry and takeovers have played an important role in accomplishing these changes; oil and gas rank first in takeover activity, with twice their proportionate share of total activity. Managers who are slow to recognize that many old practices and strategies are no longer viable are finding that takeovers are doing the job for them. Exit is cheaper to accomplish through merger and the orderly liquidation of marginal assets of the combined firms than by disorderly, expensive bankruptcy in an industry saddled with overcapacity. The end of the competitive struggle in such an industry often comes in the bankruptcy courts, with the unnecessary destruction of valuable parts of organizations that could be used productively by others.

Similarly, deregulation of the financial services market is consistent with the number 2 rank of banking/finance and the number 3 rank of insurance in Table 1. Deregulation has also been important in the transportation and broadcasting industries. Mining and minerals have been subject to many of the same forces impinging on the energy industry, including the changes in the value of the dollar.

Table 1

Intensity of industry takeover activity as measured by the value of merger and acquisition transactions in the period 1981–84 (as a percent of total takeover transactions for which valuation data are publicly reported) compared to industry size (as measured by the fraction of overall corporate market value)

Industry classification of seller	*Percent of total takeover activity*[a]	*Percent of total corporate market value*[b]
Oil and gas	26.3%	13.5%
Banking and finance	8.8	6.4
Insurance	5.9	2.9
Food processing	4.6	4.4
Mining and minerals	4.4	1.5
Conglomerate	4.4	3.2
Retail	3.6	5.2
Transportation	2.4	2.7
Leisure and entertainment	2.3	.9
Broadcasting	2.3	.7
Other	39.4	58.5

[a] Grimm, 1984, p. 41.

[b] As of 12/31/84. Total value is measured as the sum of the market value of common equity for 4,305 companies, including 1,501 companies on the NYSE, 724 companies on the ASE, plus 2,080 companies in the over-the-counter market (*The Media General Financial Weekly*, December 31, 1984, p. 17).

The development of innovative financing vehicles, such as high-yield non-investment-grade bonds (junk bonds), has removed size as a significant impediment to competition in the market for corporate control. A 1987 update by the Investor Responsibility Research Center of an earlier SEC study finds that the investment grade and high yield debt issues combined were associated with 9.8 percent of all tender offer financing from January 1981 through September 1986. Even though not yet widely used in takeovers, these new financing techniques have had important effects because they permit small firms to obtain resources for acquisition of much larger firms by issuing claims on the value of the venture (that is, the target firm's assets) just as in any other corporate investment activity.

Managerial Myopia vs. Market Myopia

It has been argued that far from pushing managers to undertake needed structural changes, growing institutional equity holdings and the fear of takeover cause managers to behave myopically and therefore to sacrifice long-term benefits to increase short-term profits. The arguments tend to confuse two separate issues: 1) whether managers are shortsighted and make decisions that undervalue future cash flows while overvaluing current cash flows (myopic managers); and 2) whether security markets are shortsighted and undervalue future cash flows while overvaluing near-term cash flows (myopic markets).

There is little formal evidence on the myopic managers issue, but I believe this phenomenon does occur. Sometimes it occurs when managers hold little stock in their companies and are compensated in ways that motivate them to take actions to increase accounting earnings rather than the value of the firm. It also occurs when managers make mistakes because they do not understand the forces that determine stock values.

There is much evidence inconsistent with the myopic markets view and no evidence that indicates it is true.

First, the mere fact that price-earnings ratios differ widely among securities indicates the market is valuing something other than current earnings. For example, it values growth as well. Indeed, the essence of a growth stock is one that has large investment projects yielding few short-term cash flows but high future earnings and cash flows. The continuing marketability of new issues for start-up companies with little record of current earnings, the Genentechs of the world, is also inconsistent with the notion that the market does not value future earnings.

Second, McConnell and Muscarella (1985) provide evidence that (except in the oil industry) stock prices respond positively to announcements of increased investment expenditures and negatively to reduced expenditures. Their evidence is also inconsistent with the notion that the equity market is myopic, since it indicates the market values spending current resources on projects which promise returns in the future.

Third, the vast evidence on efficient markets indicating that current stock prices appropriately incorporate all currently available public information is also inconsistent with the myopic markets hypothesis. Although the evidence is not literally 100 percent in support of the efficient market hypothesis, no proposition in any of the sciences is better documented.[3]

The large positive stock price reactions to announced restructurings in the oil industry are inconsistent with the notion that the market values only short-term earnings, because the restructurings involve large write-offs that reduce accounting earnings in the year. ARCO's stock price, for example, increased by 30 percent when it announced its major restructuring in 1985. The market responded positively even though ARCO simultaneously announced a $1.2 billion write-off.

Fourth, recent versions of the myopic markets hypothesis emphasize increases in the amount of institutional holdings and the pressures they face to generate high returns on a quarter-to-quarter basis. It is argued that these pressures on institutions are a major cause of pressures on corporations to generate high current earnings on a quarter-to-quarter basis. The institutional pressures are said to lead to increased takeovers of firms (because institutions are not loyal shareholders) and to decreased research and development expenditures. It is hypothesized that because R&D expenditures reduce current earnings, firms making them are therefore more likely to be taken over, and that reductions in R&D are leading to a fundamental weakening of the corporate sector of the economy.

---

[3] For an introduction to the literature and empirical evidence on the theory of efficient markets, see Elton and Gruber (1984), Chapter 15, p. 375ff. and the 167 studies referenced in the bibliography. For some anomalous evidence on market efficiency see Jensen (1978), Shiller (1981a, b). Merton (1985) provides an excellent discussion of the current state of the efficient market hypothesis.

A study of 324 firms by the Office of the Chief Economist of the SEC (April 1985) finds substantial evidence that is inconsistent with this version of the myopic markets argument. The evidence indicates the following: increased institutional stock holdings are not associated with increased takeovers of firms; increased institutional holdings are not associated with decreases in research and development expenditures; firms with high R&D expenditures are not more vulnerable to takeovers; stock prices respond positively to announcements of such increases in R&D expenditures.

Moreover, total spending on R&D is increasing concurrent with the wave of merger and acquisition activity. Total spending on R&D in 1984, a year of record acquisition activity, increased by 14 percent according to *Business Week*'s annual survey. This represented "the biggest gain since R&D spending began a steady climb in the late 1970's." All industries in the survey increased R&D spending with the exception of steel. In addition, R&D spending increased from 2 percent of sales, where it had been for five years, to 2.9 percent. In 1985 and 1986, two more record years for acquisition activity, R&D also set new records. R&D spending increased by 10 percent (to 3.1 percent of sales) in 1985, and in 1986, R&D spending again increased by 10 percent to $51 billion (to 3.5 percent of sales), in a year when total sales decreased by 1 percent.[4]

Bronwyn Hall (1987), in a detailed study of all U.S. manufacturing firms in the years 1976–85, finds in approximately 600 acquisitions that firms which are acquired do not have higher R&D expenditures (measured by the ratio of R&D to sales) than firms in the same industry which are not acquired. Also, she finds that "firms involved in mergers showed no difference in their pre- and post-merger R&D performance over those not so involved."

I know of no evidence that supports that argument that takeovers reduce R&D expenditures, even though this is a prominent argument among many of those who favor restrictions on takeovers.

A simple alternative hypothesis explains the current facts, including the criticisms of managers, quite well. Instead of supposing that the myopic market is punishing managers for their foresightedness and for being right, suppose some managers are simply mistaken—that is, their strategies are wrong— and that the financial markets are telling them they are wrong. If they don't change, their stock prices will remain low. If the managers are indeed wrong, it is desirable for the stockholders and for the economy to remove them to make way for a change in strategy and more efficient use of the resources.

The internal control mechanisms of corporations, operating through the board of directors, should encourage reluctant managers to restructure. But when the internal processes for change in large corporations are too slow, costly, and clumsy to bring about the required restructuring or change in managers efficiently, the capital markets, through the market for corporate control, are doing so. The takeover market serves as an important source of protection for investors in these situations. Other management teams that recognize an opportunity to reorganize or redeploy an organization's assets and thereby create new value can bid for the control rights in the takeover market. To be successful, such bids must be at a premium over current market value. This gives investors an opportunity to realize part of the gains from reorganization and redeployment of the assets.

---

[4] The "R&D Scoreboard" is an annual survey covering companies that account for 95 percent of total private-sector R&D expenditures. The three years referenced here can be found under "R&D Scoreboard" (1985, 1986, 1987) in the reference list. In 1984 the survey covered 820 companies; in 1985, it covered 844 companies; in 1986, it covered 859 companies.

Free Cash Flow Theory

More than a dozen separate forces drive takeover activity, including such factors as deregulation, synergies, economies of scale and scope, taxes, the level of managerial competence, and increasing globalization of U.S. markets (Roll, forthcoming). One major cause of takeover activity, the agency costs associated with conflicts between managers and shareholders over the payout of free cash flow, has received relatively little attention. Yet it has played an important role in acquisitions over the last decade.[5]

Managers are the agents of shareholders, and because both parties are self-interested, there are serious conflicts between them over the choice of the best corporate strategy. Agency costs are the total costs that arise in such arrangements. They consist of the costs of monitoring and bonding managerial behavior (such as the costs of producing audited financial statements and devising and implementing compensation plans that reward managers for actions that increase investors' wealth) and the efficiency losses that are incurred because the conflicts of interest can never be resolved perfectly. When these costs are large, the threat or actuality of takeovers can reduce them.

Free cash flow is cash flow in excess of that required to fund all of a firm's projects that have positive net present values when discounted at the relevant cost of capital. Such free cash flow must be paid out to shareholders if the firm is to be efficient and to maximize value for shareholders.

However, payment of cash to shareholders reduces the resources controlled by managers, thereby reducing the power of managers and potentially subjecting them to the monitoring by capital markets that occurs when a firm must obtain new capital. Further, managers have incentives to expand their firms beyond the size that maximizes shareholder wealth.[6] Growth increases managers' power by increasing the resources under their control, and changes in management compensation are positively related to growth.[7] Moreover, the tendency of firms to reward middle managers through promotion rather than year-to-year bonuses also creates an organizational bias toward growth to supply the new positions that such promotion-based reward systems require (Baker, 1986).

---

[5] This discussion is based on Jensen (1986).

[6] Gordon Donaldson (1984), in a detailed study of twelve large Fortune 500 firms, concludes that managers of these firms were not driven by maximization of the value of the firm, but rather by the maximization of "corporate wealth)." He defines corporate wealth (p. 3, emphasis in original) as "*the aggregate purchasing power available to management for strategic purposes during any given planning period.* ... this wealth consists of the stocks and flows of cash and cash equivalents (primarily credit) that management can use at its discretion to implement decisions involving the control of goods and services." He continues (p. 22), "In practical terms it is cash, credit, and other corporate purchasing power by which management commands goods and services."

Conflicts of interest between shareholders and managers over payout policies are especially severe when the organization generates substantial free cash flow. The problem is how to motivate managers to disgorge the cash rather than invest it at below the cost of capital or waste it through organizational inefficiencies.

The theory developed here offers a seeming paradox. Increases in financial flexibility that give managers control over free cash flow may actually cause the value of the firm to decline. This result occurs because it is difficult to assure that managers will use their discretion over resources to further the interests of shareholders.

The theory explains: (1) how debt for stock exchanges reduces the organizational inefficiencies fostered by substantial free cash flow; (2) how debt can substitute for dividends; (3) why "diversification" programs are more likely to be associated with losses than are expansion programs in the same line of business; (4) why mergers within an industry and liquidation-motivated takeovers will generally create larger gains than cross-industry mergers; (5) why the factors stimulating takeovers in such diverse businesses as broadcasting, tobacco, cable systems, and oil are essentially identical; and (6) why bidders and some targets tend to show abnormally good performance prior to takeover.

The Role of Debt in Motivating Organizational Efficiency

The agency costs of debt have been widely discussed (Jensen and Meckling, 1976; Smith and Warner, 1979), but, with the exception of Grossman and Hart (1980), the benefits of debt in motivating managers and their organizations to be efficient have largely been ignored.

Debt creation, *without retention of the proceeds of the issue*, enables managers effectively to bond their promise to pay out future cash flows. Thus, debt can be an effective substitute for dividends, something not generally recognized in the corporate finance literature.[8] Debt reduces the agency cost of free cash flow by reducing the cash flow available for spending at the discretion of managers. By issuing debt in exchange for stock, managers bond their promise to pay out future cash flows in a way that simple dividend increases do not. In doing so, they give shareholder-recipients of the debt the right to take the firm into bankruptcy court if they do not keep their promise to make the interest and principal payments.

[7] Where growth is measured by increases in sales (Murphy, 1985). This positive relationship between compensation and sales growth need not imply, although it is consistent with, causality.

[8] Literally, principal and interest payments are substitutes for dividends. Dividends and debt are not perfect substitutes, however, because interest is tax-deductible at the corporate level and dividends are not.

Of course, managers can also promise to pay out future cash flows by announcing a "permanent" increase in the dividend.[9] But because there is no contractual obligation to make the promised dividend payments, such promises are weak.[10] The fact that capital markets punish dividend cuts with large stock price reductions (Charest, 1978; Aharony and Swary, 1980) can be interpreted as an equilibrium market response to the agency costs of free cash flow.

Issuing large amounts of debt to buy back stock sets up organizational incentives to motivate managers to pay out free cash flow. In addition, the exchange of debt for stock helps managers overcome the normal organizational resistance to retrenchment that the payout of free cash flow often requires. The threat of failure to make debt-service payments serves as a strong motivating force to make such organizations more efficient.

Increased leverage also has costs. As leverage increases, the usual agency costs of debt, including bankruptcy costs, rise. The incentives to take on projects that reduce total firm value but benefit shareholders through a transfer of wealth from bondholders is one source of these costs. These costs put a limit on the desirable level of debt. The optimal debt/equity ratio is the point at which firm value is maximized, the point where the marginal costs of debt just offset the marginal benefits.

The debt created in a hostile takeover (or takeover defense) of a firm suffering severe agency costs of free cash flow need not be permanent. Indeed, sometimes "overleveraging" such a firm is desirable. In these situations, levering the firm so highly that it cannot continue to exist in its old form creates the crisis to motivate cuts in expansion programs and the sale of those divisions that are more valuable outside the firm. The proceeds are used to reduce debt to a more normal or permanent level. This process results in a reexamination of an organization's strategy and structure. When it is successful, a much leaner, more efficient, and competitive organization results.

This control hypothesis does not imply that debt issues will always have positive control effects. For example, these control effects will not be as important for rapidly growing organizations with large and highly profitable investment projects but no free cash flow. Such organizations will have to go regularly to the financial markets to obtain capital. At these times the markets have an opportunity to evaluate the company, its management, and its proposed projects. Investment bankers and analysts play an important role in this monitoring, and the market's assessment is made evident by the price investors pay for the financial claims.

---

[9] Rozeff (1982) and Easterbrook (1984a) argue that regular dividend payments can be effective in reducing agency costs with managers by assuring that managers are forced more frequently to subject themselves and their policies to the discipline of the capital markets when they acquire capital.

[10] Interestingly, Graham and Dodd (1951, Chapters 32, 34, and 36) place great importance on the dividend payout in their famous valuation formula $V = M(D + .33E)$. V is value, M is the earnings multiplier when the dividend payout rate is a "normal two-thirds of earnings," D is the expected dividend, and E is expected earnings. In their formula, dividends are valued at three times the rate of retained earnings—a proposition that has puzzled many students of modern finance (at least of my vintage). The agency cost of free cash flow that leads to over-retention and waste of shareholder resources is consistent with the deep suspicion with which Graham and Dodd viewed the lack of payout. Their discussion (chapter 34) reflects a belief in the tenuous nature of the future benefits of such retention. Although they do not couch the issues in terms of the conflict between managers and shareholders, the free cash flow theory explicated here implies that their beliefs, sometimes characterized as a conviction that "a bird in the hand is worth two in the bush," were perhaps well founded.

The control function of debt is more important in organizations that generate large cash flows but have low growth prospects, and it is even more important in organizations that must shrink. In these organizations the pressure to waste cash flows by investing them in uneconomic projects is most serious.

Leveraged Buyouts and Free Cash Flow Theory

Many of the benefits in going-private and leveraged buyout transactions seem to be due to the control function of debt. These transactions are creating a new organizational form that competes successfully with the open corporate form because of advantages in controlling the agency costs of free cash flow. In 1985, going-private and LBO transactions totaled $37.4 billion and represented 32 percent of the value of all public acquisitions. In 1986, the total value increased to $44.3 billion representing 39 percent of all public acquisitions (Baker, 1986; Grimm, 1986). Average premiums paid for publicly held firms have exceeded 50 percent.

Desirable leveraged buyout candidates are frequently firms or divisions of larger firms that have stable business histories, low growth prospects and high potential for generating cash flows; that is, situations where agency costs of free cash flows are likely to be high.

Leveraged buyouts are frequently financed with high debt; 10:1 ratios of debt to equity are not uncommon, and they average 5.25:1 (Schipper and Smith, 1986; Kaplan, 1987; DeAngelo and DeAngelo, 1986). Moreover, the use of "strip financing" and the allocation of equity in the deals reveal a sensitivity to incentives, conflicts of interest, and bankruptcy costs. Strip financing, the practice in which risky nonequity securities are held in approximately equal proportions, limits the conflict of interest among such securityholders and therefore limits bankruptcy costs. Top managers and the sponsoring venture capitalists hold disproportionate amounts of equity.

A somewhat oversimplified example illustrates the organizational effects of strip financing. Consider two firms identical in every respect except financing. Firm A is entirely financed with equity, and Firm B is highly leveraged with senior debt, subordinated debt, convertible debt, and preferred as well as common equity. Suppose Firm B securities are sold only in strips; that is, a buyer purchasing a certain percent of any security must purchase the same percent of all securities, and the securities are "stapled" together so they cannot be separated later. Security holders of both firms have identical unlevered claims on the cash flow distribution, but organizationally the two firms are very different. If Firm B managers withhold dividends to invest in value-reducing projects or if they are incompetent, stripholders have recourse to remedial powers not available to the equityholders of Firm A. Each Firm B security specifies the rights its holder has in the event of default on its dividend or coupon payment—for example, the right to take the firm into bankruptcy or to have board representation. As each security above equity goes into default, the stripholder receives new rights to intercede in the organization. As a result, it is quicker and less expensive to replace managers in Firm B.

Moreover, because every securityholder in the highly levered Firm B has the same claim on the firm, there are no conflicts between senior and junior claimants over reorganization of the claims in the event of default; to the stripholders it is a matter of moving funds from one pocket to another. Thus, Firm B will not go into bankruptcy; a required reorganization can be accomplished voluntarily, quickly, and with less expense and disruption than through bankruptcy proceedings.

Securities commonly subject to strip practices are often called "mezzanine" financing and include securities with priority superior to common stock yet subordinate to senior debt. This arrangement seems to be sensible, because several factors ignored in our simplified example imply that strictly proportional holdings of all securities is not desirable. For example, IRS restrictions deny tax deductibility of debt interest in such situations and bank holdings of equity are restricted by regulation. Riskless senior debt need not be in the strip because there are no conflicts with other claimants in the event of reorganization when there is no probability of default on its payments.

Furthermore, it is advantageous to have top level managers and venture capitalists who promote the transactions hold a larger share of the equity. Top level managers on average receive over 30 percent of the equity, and venture capitalists and the funds they represent generally retain the major share of the remainder (Schipper and Smith, 1986; Kaplan, 1987). The venture capitalists control the board of directors (in fact, they often *are* the board) and monitor managers directly. Large equity claims by managers and venture capitalists give them a strong interest in making the venture successful because their equity interests are subordinate to other claims.

Leveraged buyouts increased dramatically in the last decade from $1.2 billion in 1979, when W. T. Grimm began collecting the data, to $44.3 billion in 1986. Less than a handful of these management buyouts have ended in bankruptcy, although more have gone through private reorganizations. A thorough test of this organizational form requires the passage of time and recessions.

Some have asserted that managers engaging in a buyout of their firm are insulating themselves from monitoring. The opposite is true in the typical leveraged buyout because the venture capitalist is generally the largest stockholder and controls the board of directors. The venture capitalist therefore has both greater ability and greater incentive to monitor managers than do directors with little or no equity who represent diffuse shareholders in the typical public corporation.

Applying Free Cash Flow Theory to Takeovers

Free cash flow theory is consistent with a wide range of previously unexplained data. Here I sketch some empirical predictions of the free cash flow theory for takeovers and mergers and some of the facts that lend it credence.

The Oil Industry

The importance of takeovers and the relevance of free cash flow theory in motivating change and efficiency are particularly clear in the oil industry. Radical changes in the energy market from 1973 to the late 1970s meant that a major restructuring of the petroleum industry had to occur. The optimal level of refining and distribution capacity and crude reserves fell over this period; as of the late 1970s, the industry was plagued with excess capacity, although this was not generally recognized at the time. Reserves are reduced by reducing the level of exploration and development, and it pays to concentrate these reductions in high-cost areas such as the United States.

Substantial reductions in exploration and development and in refining and distribution capacity meant that some firms had to leave the industry. This is especially true because holding reserves is subject to economies of scale, while exploration and development are subject to diseconomies of scale.

At the same time price increases generated large cash flows, creating a particularly puzzling period in the oil industry because at the same time that change in the environment required a reduction of capacity, cash flows and profits were high; 1984 cash flows of the ten largest oil companies were $48.5 billion, 28 percent of the total cash flows of the top 200 firms in *Dun's Business Month* (July 1985) survey. This condition, in which high profits coincided with the necessity to shrink the industry, is somewhat unusual. It was caused by an increase in the average productivity of resources in the industry while the marginal productivity decreased.[11] However, management did not pay out the excess resources to shareholders. Instead, the industry continued to spend heavily on exploration and development even though the returns on these expenditures were below the cost of capital.

Paradoxically, the profitability of oil exploration and drilling activity can decrease even though the price of oil increases, if the value of reserves in the ground falls. This decrease can occur when the price increase is associated with reductions in consumption that make marketing newly discovered oil difficult. In the late 1970s the increased holding costs associated with higher real interest rates, reductions in expected future oil price increases, increased exploration and development costs, and contrived reductions in current supply (and thus larger future potential supply) combined to make many current exploration and development projects uneconomic. The industry, however, continued to spend heavily on such projects.

The waste associated with excessive exploration and development expenditures explains why buying oil on Wall Street was considerably cheaper than obtaining it by drilling holes in the ground, even after adjustment for differential taxes and regulations on prices of old oil. Wall Street was not undervaluing the oil; it was valuing it correctly, but it was also correctly valuing the wasted expenditures on exploration and development that oil companies were making. When these managerially imposed "taxes" on the reserves were taken into account in stock prices, the net prices of oil on Wall Street was low. This low price provided incentives for firms to obtain reserves by purchasing other oil companies and reducing expenditures on non-cost-effective exploration. In this way the capital markets provided incentives for firms to make adjustments that were not effectively motivated by competition in the product markets.

The fact that oil industry managers tried to invest funds outside the industry is also evidence that they could not find enough profitable projects within the industry to use the huge inflow of resources efficiently. Unfortunately these efforts failed. The diversification programs involved purchases of companies in retailing (Marcor by Mobil), manufacturing (Reliance Electric by Exxon), office equipment (Vydec by Exxon), and mining (Kennecott by Sohio, Anaconda Minerals by ARCO, Cyprus Mines by Amoco). These acquisitions turned out to be among the least successful of the last decade, partly because of bad luck (like the collapse of the minerals industry) and partly because of a lack of managerial expertise outside the oil industry. In sum, the stage was set for retrenchment in the oil industry in the early 1980s. Yet the product and capital markets could not force management to change its strategy because the industry's high internal cash flows insulated them from these pressures.

---

[11] More detailed analysis of this point is available in Jensen (1987).

Ultimately the capital markets, through the takeover market, forced managers to respond to the new market conditions. T. Boone Pickens of Mesa Petroleum perceived early that the industry had to be restructured. Partly as a result of Mesa's efforts, firms in the industry were led to merge, and in the merging process they paid out large amounts of capital to shareholders, reduced excess expenditures on exploration and development, and reduced excess capacity in refining and distribution. The result has been large gains in efficiency. Total gains to shareholders in the Gulf/Chevron, Getty/Texaco and Du Pont/Conoco mergers, for example, were over $17 billion. Much more is possible. Jacobs (1986) estimates total potential gains of approximately $200 billion from eliminating the inefficiencies in 98 petroleum firms as of December 1984.

Recent events indicate that actual takeover is not necessary to induce the required adjustments; the Phillips, Unocal and Arco restructurings all involve large stock repurchases with debt and cash, increases in dividend payments, and reductions in exploration and development. They generated increases of 20 percent to 35 percent in market value, totaling $6.6 billion.

Other Industries in Theory and Practice

Acquisitions are one way managers spend cash instead of paying it out to shareholders. Free cash flow theory implies that managers of firms with unused borrowing power and large free cash flows are more likely to undertake low-benefit or even value-destroying mergers. Diversification programs generally fit this category, and the theory predicts that they will generate lower total gains. Thus, some acquisitions are a solution to the agency problem of free cash flow while others, such as diversification programs, are symptoms of those problems.

The major benefit of diversification mergers may be that they involve less waste of resources than if the funds had been invested internally in unprofitable projects. Acquisitions made with cash or securities other than stock involve payout of resources to shareholders of the target company, and this can create net benefits even if the merger creates operating inefficiencies. To illustrate, consider an acquiring firm with substantial free cash flow that the market expects will be invested in low-return projects with a negative net present value of $100 million. If this firm uses up its free cash flow (and thereby prevents its waste) by acquiring another firm that generates zero synergies, the combined market value of the two firms will rise by $100 million. The market value increases because the acquisition eliminates the expenditures on internal investments with negative market value of $100 million.

Because the bidding firms are using funds that would otherwise have been spent on low or negative-return projects, the opportunity cost of the funds is lower than their cost of capital. As a result, they will tend to overpay for the acquisition and thereby transfer some, if not all, of the gains to the target firm's shareholders. In extreme cases they may pay so much that the bidding firm's share price falls, in effect giving the target shareholders more than 100 percent of the gains. These predictions are consistent with the evidence that shareholders of target companies reap most of the gains from a takeover.

Low-return mergers are more likely to occur in industries with large cash flows whose economics dictate retrenchment. Horizontal mergers (where cash or debt is the form of payment) within declining industries will tend to create value because they facilitate exit—the cash or debt payments to shareholders of the target firm cause resources to leave the industry directly. When Socal acquired Gulf in 1984 for $13.2 billion in cash, the oil industry shrank by $13.2 billion as soon as the checks were mailed. Mergers outside the declining industry are more likely to have low or even negative returns because managers are likely to know less about managing such firms. Oil fits this description, and so does tobacco. Tobacco firms face declining demand as a result of changing smoking habits but generate large free cash flow and have been involved in major diversifying acquisitions, as in the $5.6 billion purchase of General Foods by Philip Morris. The theory predicts that these acquisitions in nonrelated industries are more likely to reduce productivity, although the positive total gains to buyers and sellers indicate these negative productivity effects are outweighed by the reductions in waste from internal expansion.

Forest products is another industry with excess capacity and acquisition activity, including the acquisition of St. Regis by Champion International and Crown Zellerbach by Sir James Goldsmith. Horizontal mergers for cash or debt in such an industry generate gains by encouraging exit of resources (through payout) and by substituting existing capacity for investment in new facilities by firms that are short of capacity. Food industry mergers also appear to reflect the expenditure of free cash flow. The industry apparently generates large cash flows with few growth opportunities. It is, therefore, a good candidate for leveraged buy-outs, and these are now occurring; the $6.3 billion Beatrice LBO is the largest ever.

The broadcasting industry generates rents in the form of large cash flows on its licenses and also fits the theory. Regulation limits the overall supply of licenses and the number owned by a single entity. Thus, profitable internal investments are limited, and the industry's free cash flow has been spent on organizational inefficiencies and diversification programs, making these firms takeover targets. The CBS debt-for-stock exchange and restructuring as a defense against the hostile bid by Turner fits the theory, and so does the $3.5 billion purchase of American Broadcasting Company by Capital Cities Communications. Complete cable systems also create agency problems from free cash flows in the form of rents on their franchises and quasi rents on their installed capital and are likely to generate free cash flow problems. Drug companies with large cash flows from previous successful discoveries and few potential future prospects are also candidates for large agency costs of free cash flow.

Free cash flow theory predicts that many acquirers will tend to perform exceptionally well prior to acquisition. Empirical evidence from studies of both stock prices and accounting data indicates exceptionally good performance for acquirers prior to acquisition (Magenheim and Mueller, 1985; Bradley and Jarrell, 1985). This exceptional stock price performance is often associated with increased free cash flow, which is then used for acquisition programs as observed in the oil industry. Targets will be of two kinds: firms with poor management that have done poorly before the merger, and firms that have done exceptionally well and have large free cash flow that they refuse to pay out to shareholders. Both kinds of targets seem to exist.[12]

The theory predicts that takeovers financed with cash and debt will create larger benefits than those accomplished through exchange of stock. Stock acquisitions do nothing to take up the organizations' financial slack and are therefore unlikely to motivate managers to use resources more efficiently. The recent evidence on takeover premiums is consistent with this prediction.[13]

In the best study to date of the determinants of takeover, Palepu (1986) finds strong evidence consistent with the free cash flow theory of mergers. He studied a sample of 163 firms acquired in the period 1971–79 and a random sample of 256 firms that were not acquired. Both samples were in mining and manufacturing and were listed on either the New York or American stock exchange. He finds that firms with a mismatch between growth and resources are more likely to be taken over. These are firms with high growth (measured by average sales growth), low liquidity (measured by the ratio of liquid assets to total assets) and high leverage, and firms with low growth, high liquidity, and low leverage. He also finds that poor prior performance (measured by the net of market returns in the four years before the acquisition) is significantly related to the probability of takeover and, interestingly, that accounting measures of past performance such as return on equity are unrelated to the probability of takeover.

Free cash flow is only one of the many factors that go into a takeover decision. But the evidence indicates that it is an important factor and provides a useful perspective on the conflict.

Controversial Issues from an Agency Perspective

High-Yield, Non-Investment Grade Bonds: "Junk" Bonds

The past several years have witnessed a major innovation in the financial markets —the establishment of active markets in high-yield bonds. These bonds, rated below investment grade by the bond-rating agencies, are frequently referred to as junk bonds, a disparaging term that bears no relation to their pedigree. High-yield bonds are best viewed as commercial loans that can be resold in secondary markets. They are further evidence of the securitization that has converted formerly illiquid financial claims such as mortgages into marketable claims. Total publicly held high-yield bonds have risen from $7 billion in 1970 to $125 billion in 1986, or 23 percent of the total corporate bond market (Taggart, 1986; Drexel Burnham Lambert, 1987). By traditional standards they are more risky than investment-grade bonds and therefore carry interest rates 3 to 5 percentage points higher than the yields on government bonds of comparable maturity. In an early study, Blume and Keim (1984) find that the default rates on these bonds have been low and the realized returns have been disproportionately higher than their risk.

---

[12] Asquith (1983) finds evidence of below-normal stock price performance for 302 target firms in the 400 days before 20 days prior to the takeover bid. Mandelker (1974) finds negative abnormal performance for target firms in the period from 40 months before until 9 months before the outcome of the merger bid is known. Langtieg (1978) reports significant negative returns in the period from 72 months before until 19 months before the outcome date, but positive abnormal returns in the 19 months preceding the merger date.

[13] See Wansley, Lane and Yang (1987) who find higher returns to targets and to bidders in cash transactions, and Wansley and Fayez (1986).

High-yield bonds have been attacked by those who wish to inhibit their use, particularly in the financing of takeover bids. However, companies commonly raise funds to finance ventures by selling claims to be paid from the proceeds of the venture; this is the essence of debt or stock issues used to finance new ventures. High-yield bonds used in takeovers work similarly. The bonds provide a claim on the proceeds of the venture, using the assets and cash flows of the target plus the equity contributed by the acquirer as collateral. Similarly, individuals purchase homes using the home plus their down payment as collateral for the mortgage. The structure of this contract offers nothing inherently unusual.

Some might argue that the risk of high-yield bonds used in takeover attempts is "too high." But high-yield bonds are less risky by definition than common stock claims on the same venture, since the claims of common stockholders are subordinate to those of the holders of high-yield bonds. Would these same critics argue that the stock claims are too risky and thus should be barred? The risk argument makes logical sense only as an argument that transactions costs associated with bankruptcy or recontracting are too high in these ventures or that the bonds are priced too high and that investors who purchase them will not earn returns high enough to compensate for the risk they are incurring. This overpricing argument makes little sense because there is vast evidence that investors are capable of pricing risks in all sorts of other markets.

In January 1986 the Federal Reserve Board issued a new interpretation of the margin rules that restricts the use of debt in takeovers to 50 percent or less of the purchase price. The rule has had little effect on takeovers because bidders otherwise subject to the constraint have instead used high-yield preferred stock rated below investment grade which is converted to debt after completion of the acquisition or bridge loans. This rule was apparently motivated by the belief that the use of corporate debt has become abnormally and dangerously high and was threatening the economy. This assessment is not consistent with the facts. Table 2 presents measures of debt use by nonfinancial corporations in the United State. The debt-equity ratio is measured relative to three bases: market value of equity, estimated current asset value of equity, and accounting book value of equity measured at historical cost.

Although debt-equity ratios were higher in 1985 than in 1961, they were not at record levels, The book value debt-equity ratio reached a high of 81.4 percent in 1984

Table 2
Debt-to-equity ratios: non-financial corporations

Year	*Book value*	*Current value*	*Market value*
1961	57.1%	41.1%	38.5%
1962	58.2	42.5	45.6
1963	59.6	44.5	41.7
1964	59.9	45.4	39.8
1965	61.1	46.5	40.0
1966	62.7	47.4	48.4
1967	64.7	48.7	41.3
1968	67.2	50.5	40.2
1969	68.1	50.3	50.3
1970	70.5	50.7	54.7
1971	70.4	50.7	50.0
1972	70.2	50.3	48.1
1973	70.9	48.9	67.7
1974	70.2	43.9	105.2
1975	66.7	41.6	79.5
1976	65.6	41.1	74.2
1977	67.7	41.4	87.6
1978	69.1	41.1	94.8
1979	69.9	39.9	88.7
1980	68.3	37.8	70.0
1981	71.0	38.3	82.7
1982	74.3	40.0	77.7
1983	73.0	40.6	69.2
1984	81.4	46.1	80.5
1985	78.0	46.5	60.8

Source: Federal Reserve Board (1986)

but declined to 78 percent in 1985. Debt-equity ratios measured on a historical cost basis are relatively high because of the previous decade of inflation. Maintenance of the same inflation-adjusted debt ratios in time of inflation implies that the book value ratio must rise because the current value of assets in the denominator of the inflation-adjusted ratio is rising. The current value ratio, which takes account of inflation, fell from 50.7 percent in 1970 to 46.5 percent in 1985. The market-value ratio rose from 54.7 percent in 1970 to 80.5 percent in 1984 and plummeted to 60.8 percent in 1985. The 1985 market-value ratio was 45 percentage points below its 1974 peak of 105.2 percent. Thus, the Federal Reserve System's own data are inconsistent with the reasons given for its restrictions on the use of debt.

High-yield bonds were first used in a takeover bid in 1984 and were involved in relatively few bids in total. In 1984, only about 12 percent of the $14.3 billion of new high-yield debt was associated with mergers and acquisitions. In 1985, 26 percent of the $14.7 billion of new high-yield debt was used in acquisitions.[14] According to *Mergers & Acquisitions*, 1986 acquisition-related high-yield debt still represents less than one of every 12 dollars in acquisition value. Nevertheless, high-yield bonds are an important innovation in the takeover field because they help eliminate size as a deterrent to takeover. They have been particularly influential in helping to bring about reorganizations in the oil industry.

Historical default rates on high-yield bonds have been low, but many of the bonds are so new that the experience could prove to be different in the next downturn. Various opponents (including executives who desire protection from the takeover market and members of the financial community, such as commercial banks and insurance companies, who want to restrict competition from this new financing vehicle) have backed regulations and legislation to restrict the issuance of high-yield bonds, to penalize their tax status, and to restrict their holding by thrifts, which can now buy them as substitutes for the issuance of nonmarketable commercial loans. These proposals are premature, to say the least.

Severance Contracts: "Golden Parachutes"

The increasing sophistication of takeover experts and the availability of high-yield bond financing for profitable takeover ventures means that the largest of the Fortune 500 companies are now potentially subject to takeover; mere size is no longer an effective defense. This susceptibility to takeover has created a new contracting environment for top-level managers. Roughly 50 percent of the top level managers of target firms are gone within three years of acquisition—either hostile or voluntary. Many managers are legitimately anxious, and it will take time for the system to work out an appropriate set of practices and contracts reflecting the risks and rewards of the new environment.

Unfortunately, a major component of the solution to the conflict of interest between shareholders and managers has been vastly misunderstood. I am referring to severance contracts that compensate managers for the loss of their jobs in the event of a change in control. These have been popularly labeled "golden parachutes."

These control-related contracts are beneficial when correctly implemented, because they help reduce the conflict of interest between shareholders and managers at times of takeover and therefore make it more likely that the productive gains stemming from changes in control will be realized. The evidence indicates that stock prices of firms that adopt severance-related compensation contracts for managers on average rise about 3 percent when adoption of the contracts is announced (Lambert and Larcker, 1985). There is no easy way to tell what proportion of the effect is due to the market interpreting the announcement as a signal that a takeover bid is more likely and what proportion is due to the reduction in conflict between managers and shareholders.

---

[14] Source: Drexel Burnham Lambert, private correspondence.

At times of takeover, shareholders are implicitly asking the top-level managers of their firm to negotiate a deal for them that frequently involves the imposition of large personal costs on the managers and their families. These involve moving costs, the loss of position, power and prestige, and even the loss of their jobs. Shareholders are asking the very people who are most likely to have invested considerable time and energy (in some cases a life's work) in building a successful organization to negotiate its sale and the possible redirection of its resources.

It is important to confront these conflicts and to structure contracts with managers to reduce them. It would make no sense to hire a realtor to sell your house and then penalize him for doing so. Yet that is the implication of many of the emotional reactions to control-related severance contracts. The restrictions and tax penalties imposed on these severance payments by the Deficit Reduction Act of 1984 are unwise interferences in the contracting freedoms of shareholders and managers and should be eliminated.

Golden parachutes can be used to restrict takeovers and to entrench managers at the expense of shareholders. The key to deciding whether a contract is well-designed is whether it helps solve the conflict-of-interest problem between shareholders and managers. Solving this problem requires extending control-related severance contracts beyond the chief executive to those members of the top-level management team who must play an important role in negotiating and implementing any transfer of control. Contracts that award severance contracts to substantial numbers of managers beyond this group are unlikely to be in the shareholders' interest. Beneficial Corp. awarded such contracts to over 200 of its managers (Morrison, 1982). These are likely to be difficult to justify as in the shareholders' interests.

Severance-related compensation contracts are particularly important in situations where it is optimal for managers to invest in organization-specific human capital; that is, in skills and knowledge that have little or no value in other organizations. Managers will not so invest where the likelihood is high that their investment will be eliminated by an unexpected transfer of control and the loss of their jobs. In such situations, the firm will have to pay for all costs associated with the creation of such organization-specific human capital, and it will be more costly for the firm to attract and retain highly talented managers when they have better opportunities elsewhere. In addition, contracts that award excessive severance compensation to the appropriate group of managers will tend to motivate managers to sell the firm at too low a price.

No simple rules can be specified that will easily prevent the misuse of golden parachutes because the appropriate solution will depend on many factors that are specific to each situation (like the amount of stock held by the managers and the optimal amount of investment in organization-specific human capital). In general, contracts that award inappropriately high payments to an excessively large group will reduce efficiency and harm shareholders by raising the cost of acquisition and by transferring wealth from shareholders to managers. The generally appropriate solution is to make the control-related severance contracts pay off in a way that is tied to the premium earned by the stockholders. Stock options or restricted stock appreciation rights that pay off only in the event of a change in control are two options that have

Accession Number.
121814
Date 7:12:89

some of the appropriate properties. In general, policies that encourage increased stock ownership by managers and the board of directors will provide incentives that will tend to reduce the conflicts of interests with managers.

Targeted Repurchases: "Greenmail"

Most proposals to restrict or prohibit targeted repurchases (transactions pejoratively labeled "greenmail") are nothing more than antitakeover proposals in disguise. Greenmail is actually a targeted repurchase, an offer by management to repurchase the shares of a subset of shareholders at a premium, an offer not made to other shareholders. Greenmail is an appellation that suggests blackmail; yet the only effective threat possessed by a greenmailer is the right to offer to purchase stock from shareholders at a substantial premium. The "damage" to shareholders caused by this action is difficult to find. Those who propose to "protect" shareholders by paying greenmail hide this fact behind emotional language designed to mislead. But management can easily prohibit greenmail without legislation: it need only announce a policy that prohibits the board or management from making such payments.

The ease with which managers can prevent targeted repurchases makes it clear that the problem lies with managers who use such payments to protect themselves from competition in the market for corporate control. Three careful studies of these transactions indicate that, when measured from the initial toehold purchase to the final repurchase of the shares, the stock price of target firms rises (Holderness and Sheehan, 1985; Mikkelson and Ruback 1985, 1986). Therefore, shareholders are benefited, not harmed, by the whole sequence of events. However, when greenmail is used to buy off an acquirer who has made an offer for the firm, shareholders are harmed by the loss of the takeover premium. There is some indication that the stock price increases represent the expectation of future takeover premiums in firms in which the targeted repurchase was not sufficient to prevent ultimate takeover of the firm (see Mikkelson and Ruback, 1986). If so, then, as in the final defeat of tender offers found by Bradley, Desai and Kim (1983), all premiums are lost to shareholders in firms for which the repurchase and associated standstill agreements successfully lock up the firm, preventing any voluntary reorganization.

Problems in the Delaware Court: "Poison Pills"

Delaware courts have created over the years a highly productive fabric of corporate law that has benefited the nation. The court is having difficulty, however, sorting out the complex issues it faces in the takeover area. The result has been a confusing set of decisions that, in contrast to much of the court's previous history, appears to make little economic sense.[15]

---

[15] See, for example, *Moran v. Household Intl., Inc.*, 490 A.2d 1059 (Del. Ch. 1985) aff'd 500 A.2d 1346 (Del. 1985) (upholding poison pill rights issue), *Smith v. Van Gorkom*, 488 A.2d 858, (holding board liable for damages in sale of firm at substantial premium over market price), *Unocal v. Mesa*, 493 A.2d 946, 954 (Del. 1985) (allowing discriminatory targeted repurchase that confiscates wealth of largest shareholder), *Revlon Inc. v. MacAndrews & Forbes Holdings Inc.* 506 A.2d 173, (Del. 1986) (invalidation of Revlon's lockup sale of a prime division to Forstmann Little at a below-market price).

One key case involved a unilateral decision from the board of directors of Household International to change the nature of the contractual relationship with Household's shareholders in a fundamental way. Effectively, the board restricted the alienability of the common stock by prohibiting shareholders from selling their shares, without permission of the board, into a control transaction leading to merger at a premium over market value lower than about $6 billion. Since Household had a market value of about $2 billion at the time, this step prevented its shareholders from accepting any premium less than 200 percent—more than four times the average takeover premium of 50 percent common in recent times. This decision is difficult to justify as in the shareholders' interests, but the Delaware Supreme Court upheld in November 1985 the right of the board to take such action.

The Delaware court's model of the corporation is founded in the business judgment rule—the legal doctrine that holds that unless explicit evidence of fraud or self-dealing exists the board of directors is presumed to be acting in the interests of the shareholders. In particular, the board is presumed to act altruistically and never out of incentives to preserve the interests of managers or their own positions as board members.

The altruistic model of the board that is the implicit foundation of the business judgment rule is obviously incorrect as a description of human behavior. But in spite of its falsity, the altruistic model has been sufficiently robust to yield good law for a wide range of cases for many years. Alternative agency models of the corporation that incorporate conflicts of interest between board members and shareholders are much more complicated.[16] The court is now being forced to deal with these complexities, but it is doing so with an inadequate analytic foundation.

To illustrate the problem the Delaware court is facing, consider the simple situation in which a principal (stockholder) hires an agent (managers and board of directors) to take some actions on his or her behalf. To effect this arrangement, the principal delegates to the agent a set of decision rights. On entering the relationship, the principal wants the courts to enforce the contract that delegates decision rights to the agent. One of the primary purposes of the business judgment rule is to keep the courts out of the business of second-guessing the agent's decisions and holding the agent liable for damages. Doing so would make it difficult or impossible for principals to hire agents in the first place, which would eliminate the benefit of specialization. Under this rule, the court refused to second-guess the decision of the Household board of directors.

The principal may want to delegate a wide range of decision rights to the agent. In no event, however, will it be sensible for the principal to delegate the ultimate control rights to the agent: the rights to hire, fire, and set the compensation of the agent (Fama and Jensen, 1985). If the principal were to delegate the control rights to the agent, the agent could not be fired and would have the right to set his own compensation. In this circumstance the agent would become the effective owner of the decision rights (although he probably could not alienate them) and could be expected to use them in his own interests.

---

[16] Easterbrook (1984b) and Jensen and Smith (1985) provide summaries of much of the work in the area.

If the business judgment rule is applied to conflicts over control rights between principals and agents, the courts are effectively giving the agent the right to change the control rights unilaterally. In the long run, this interpretation of the contract will destroy the possibility of such cooperative arrangements, because it will leave principals with few effective rights.

The courts have applied the business judgment rule to conflicts between management and shareholders over the issuance of poison pill preferred stock, poison pill rights, and discriminatory targeted repurchases, and have given managers and boards the rights to use these devices.[17] Poison pill securities change fundamental aspects of the corporate rules that govern the relationship between shareholders, managers, and the board of directors when a control-related event occurs. They are called "poison pills" because they alter the company to make it indigestible to an acquirer. In doing so, the courts are essentially giving the agents (managers and the board) the right to change unilaterally critical control aspects of the contract, in particular, the right to prevent the firing of the agents.

The Delaware court decision upholding the decision of the board of Household International has unleashed a flood of poison pill adoptions by American corporations. Ryngaert (forthcoming) and Malatesta and Walkling (forthcoming) study the effects of over 300 of these plans adopted primarily in the period since the *Household* and *Unocal* decisions. They find statistically significant negative stock price effects on the announcement of the adoption of the plans; they find also that the plans tend to be implemented in firms in which managers own relatively little of their firm's stock. Malatesta and Walkling also find that firms adopting such plans are significantly less profitable than other firms in their industries in the year prior to adoption.

The court has erred in allowing the Household board, under the business judgment rule, to make the fundamental change in the structure of the organization implied by the rights issue without vote of its shareholders. Several other poison pill cases have been heard by the courts with similar outcomes, but one New Jersey and two New York courts have recently ruled against poison pills that substantially interfere with the voting rights of large-block shareholders.[18] An Illinois district court recently voided a poison pill (affirmed by the Seventh Circuit Court of Appeals) and two weeks later approved a new pill issued by the same company.[19]

The problem with these special securities and the provision they contain is not with their appropriateness (some might well be desirable), but with the manner in which they are being adopted; that is, without approval by shareholders. Boards of directors show little inclination to refer such issues to shareholders.

The continued application of the business judgment rule to conflicts over control has far-reaching consequences. If the current trend continues, this process will erode the limits to judicial interference in the management of corporations historically provided by the business judgment rule and severely cripple the corporation in the competition for survival. Indeed, the protection afforded managers by the business judgment rule is already eroding.[20] The court seems to be imposing a higher standard on corporations that adopt a poison pill. This erosion of the business judgment rule appears to be motivated by the court's understanding that the pill gives management and the board great power. So the court is brought into the business of second-guessing managers' business decisions. The court currently seems to be inclined to give this scrutiny only to control transactions.

---

[17] *Moran v. Household Intl.*, and *Unocal v. Mesa*. Full citations provided in note 15.

[18] *Ministar Acquiring corp. v. AMF Inc.*, 621 Fed Sup 1252. Dis NY, 1985 *Unilever Acquisition Corp. v. Richardson-Vicks, Inc.*, 618 Fed Supp 407. So Dist. NY 1985, *Asarco Inc. v. M.R.H. Holmes a Court*, 611 Fed Sup 468. Dist. Ct. of NJ, 1985, and *Dynamics Corp. of America v. CTS Corporation*.

[19] *Dynamics Corp. of America v. CTS Corp., et al.* U.S. District Court, Northern District of Illinois, Eastern Division, No. 86 C 1624, (April 17, 1986), affirmed Seventh Circuit Court of Appeals Nos. 86-1601, 86-1608, and *Dynamics Corp. of America v. CTS Corp., et al.* (May 3, 1986).

I believe that this erosion of the business judgment rule will be checked because the court will recognize the problems with its current approach. The easiest solution to the problem is for the court to deny protection under the business judgment rule to managerial decisions on control issues unless those decisions have been ratified by shareholder vote.[21]

SEC 13d Disclosure Rules and the Creation of Externalities

It has become popular to argue there is too much takeover activity. Yet the opposite is most likely true because of free-riding problems caused by the current regulations that require disclosure of holdings and intentions of the purchaser in SEC 13d reports.[22] These reports must be filed within 10 days of acquisition of 5 percent or more of a company's shares and must disclose the number of shares owned, the identity of the owner, and the purpose of the acquisition. Current rules allow the acquiring firm to buy as many additional shares as it can in the 10-day window between the time the 5 percent filing barrier is reached and the time of filing. This rule allows buyers to acquire shares that average 13.9 percent of the target firm.[23]

Since market prices adjust to the expected value of the takeover bid immediately after the 13d announcement, most of the acquirer's profits are made almost entirely on the difference between the price paid for the shares purchased prior to the filing of the 13d and their value after the acquisition. This drives a wedge, however, between the private benefits earned by the acquirer and the total social benefits of the acquisition; the acquirer pays 100 percent of the acquisition costs and, on average, captures less than 14 percent of the benefits. The remaining benefits go to the other shareholders. The activities of Mesa Petroleum, for example, have yielded benefits to the shareholders of companies involved in its transactions that exceed $13 billion. Mesa itself has paid hundreds of millions of dollars in financing, legal, and investment banking fees and borne all the risks of loss. Yet it has earned only about $750 million on these transactions.

---

[20] See *Revlon Inc. v. MacAndrews & Forbes Holdings Inc.*, 506 A.2nd 173, 180 (Del. 1986), in which the court seems to be reviewing very detailed aspects of the board's decision leading to the invalidation of Revlon's lockup sale of a prime division to Forstmann Little at a below-market price. The Court of Appeals for the Second Circuit in *Hanson Trust v. SCM Corp.*, (Nos. 85-7951, 85-7953, 2d Cir. Jan. 6, 1986) (written opinion filed Jan. 6, 1986), enjoined lockups given by SCM defending itself from takeover by Hanson Trust. See also Herzel, Colling, and Carlson (1986) for detailed analysis of these cases and lockups in general.

[21] See Koleman (1985) and Investor Responsibility Research Center (1985).

[22] Grossman and Hart (1980) present an extensive discussion of the free-riding problem in corporate takeovers.

[23] Mikkelson and Ruback (1985) show that average holdings for 397 initial 13d filings in 1928–80 is 20.9 percent. However, 120 of these 13ds (with average holdings of 37 percent) were filed simultaneously with announcement of a takeover. Eliminating these yields estimated average holdings for non-takeover filings of 13.9 percent.

Consider an acquisition that promises total expected gains of $100 million. If the acquirer expects to capture only $14 million of this amount if the bid is successful, the bid will occur only if the legal, investment banking, and other costs (including the required risk premium) are less than $14 million. All such acquisitions that are expected to cost more than this will not be made, and shareholders and society are thus denied the benefits of those reorganizations. If the costs, for example, are expected to be $15 million, the bid will not occur and the $85 million benefit will not be realized.

The solution to this problem is to abolish the SEC 13d reporting requirement or to increase significantly the trigger point from the current 5 percent level. Unfortunately, current proposals in Congress to reduce the 10-day window to one or two days and to reduce the trigger point to 1 percent or 2 percent are moves in exactly the wrong direction. It is clear why antitakeover forces want such restrictive legislation. But the effect of this proposal is equivalent to that of an anti-patent law which requires public disclosure of all inventions and denies the inventor all but a one or two percent property right in the proceeds of his or her invention. Shareholders will clearly be harmed by such regulation.

Conclusion

Although economic analysis and the evidence indicate that the market for corporate control is benefiting shareholders, society, and the corporation as an organizational form, it is also making life more uncomfortable for top level executives. This discomfort is creating strong pressures at both the state and federal levels for restrictions that will seriously cripple the working of this market. In 1985, 1986 and 1987 there were dozens of bills in the congressional hopper proposing various restrictions on the market for corporate control. Others proposed major new restrictions on share ownership and financial instruments. Within the past several years the legislatures of numerous states have passed antitakeover laws and the Supreme Court has recently upheld the Indiana law that prohibits someone who purchases 20 percent or more of a firm's shares without permission of the board of directors from voting those shares unless such approval is granted by a majority vote of disinterested shareholders. The New York state law bars the purchaser of even 100 percent of a firm's shares from doing anything with the assets for five years unless permission of the incumbent board is obtained.

This political activity is another example of special interests using the democratic political system to change the rules of the game to benefit themselves at the expense of society as a whole. In this case, the special interests are top level corporate managers and other groups who stand to lose from competition in the market for corporate control. If these special interests are successful, the results will be a reduction in efficiency and a significant weakening of the corporation as an organizational form.

■ *This research is supported by the Division of Research, Harvard Business School, and the Managerial Economics Research Center, University of Rochester. I am grateful for the research assistance of Michael Stevenson and the helpful comments of Sidney Davidson, Harry DeAngelo, Jay Light, Robert Kaplan, Nancy Macmillan, David Mullins, Susan Rose-Ackerman, Richard Ruback, Carl Shapiro, Timothy Taylor, Wolf Weinhold, Toni Wolcott, and, especially, Armen Alchian. The analysis here draws heavily on that in Jensen (forthcoming).*

References

Aharony, Joseph, and Itzhak Swary, "Quarterly Dividend and Earnings Announcements and Stockholder's Returns: An Empirical Analysis," *Journal of Finance*, 1980, *35*, 1-12.

Asquith, Paul R., "Merger Bids, Uncertainty, and Stockholders Returns," *Journal of Financial Economics*, 1983, *11*, 51-83.

Asquith, Paul R., Robert F. Bruner, and David Mullins, "Merger Returns and the Form of Financing," unpublished working paper, Harvard Business School, June 1987.

Baker, George, "Compensation and Hierarchies," Harvard Business School, January, 1986.

Blume, Marshall E., and Donald B. Keim, "Risk and Return Characteristics of Lower-Grade Bonds," The Wharton School of Finance, December 1984.

Bradley, Michael, Anand Desai, and E. Han Kim, "The Rationale Behind Interfirm Tender Offers: Information or Synergy? *Journal of Financial Economics*, April 1983, *11*, 183-206.

Bradley, Michael, and Gregg Jarrell, "Evidence on Gains from Mergers and Takeovers," presented at the Conference on Takeovers and Contests for Corporate Control, Columbia University, November 1985.

Bruner, Robert F., "The Use of Excess Cash and Debt Capacity as a Motive for Merger," Colgate Darden Graduate School of Business, December 1985.

"Cash Flow: The Top 200," *Dun's Business Month*, July 1985, 44-50.

Charest, Guy, "Dividend Information, Stock Returns, and Market Efficiency—II," *Journal of Financial Economics*, 1978, *6*, 297-330.

DeAngelo, Harry, and Linda DeAngelo, "Management Buyouts of Publicly Traded Corporations," In Copeland, Thomas E., ed., *Modern Finance and Industrial Economics: Papers in Honor of J. Fred Weston*. Oxford, England: Basil Blackwell, 1986.

Donaldson, Gordon, *Managing Corporate Wealth*. New York: Praeger, 1984.

Drexel Burnham Lambert, "The Case for High Yield Bonds," company report, 1987, 4-5.

Easterbrook, F. H., "Two Agency-Cost Explanations of Dividends," *American Economic Review*, 1984a, *74*, 650-59.

Easterbrook, F. H., "Managers' Discretion and Investors' Welfare: Theories and Evidence," *Delaware Journal of Corporate Law*, 1984b, *9*, No. 3, 540-571.

Elton, E., and M. Gruber, *Modern Portfolio Theory and Investment Analysis*. New York: Wiley, 1984.

Fama, Eugene F., and Michael C. Jensen, "Organizational Forms and Investment Decisions," *Journal of Law and Economics*, 1985, *14*, 101-119.

Federal Reserve Board, "Balance Sheets, Flow of Funds." Washington, D.C.: GPO, October 1986.

Graham, Benjamin, and David L. Dodd, *Security Analysis: Principles and Technique*. New York: McGraw-Hill, 1951.

Grimm, W. T., *Mergerstat Review*. Annual editions 1984, 1985, 1986.

Grossman, S., and Oliver Hart, "Takeover Bids, the Free-Rider Problem, and the Theory of the Corporation," *Bell Journal of Economics*, Spring 1980, 42-64.

Herzel, Leo, Dale E. Colling, and James B. Carlson, "Misunderstanding Lockups," *Securities Regulation*, September 1986, 150–180.

Hall, Bronwyn, "The Effect of Takeover Activity on Corporate Research and Development," presented at NBER Conference on the Effects of Mergers & Acquisitions, February 1987.

Holderness, Clifford G., and Dennis P. Sheehan, "Raiders or Saviors? The Evidence of Six Controversial Investors," *Journal of Financial Economics*, December 1985, *14*, 555–579.

Investor Responsibility Research Center, Inc., "Voting by Institutional Investors on Corporate Governance Questions," 1985 Proxy Season, Corporate Governance Service, 19–25.

Jacobs, E. Allen, "The Agency Cost of Corporate Control: The Petroleum Industry," Massachusetts Institute of Technology, March 1986.

Jensen, Michael C., ed., "Symposium on Some Anomalous Evidence on Market Efficiency," *Journal of Financial Economics*, June/September 1978, *6*, 95–101.

Jensen, Michael C., "Takeovers: Folklore and Science," *Harvard Business Review*, Nov.-Dec. 1984, 109–121.

Jensen, Michael C., "Agency Costs of Free Cash Flow, Corporate Finance and Takeovers," *American Economic Review*, May 1986, *76*, 323–29.

Jensen, Michael C., "The Takeover Controversy: Analysis and Evidence." In Coffee, John, Louis Lowenstein, and Susan Rose-Ackerman, eds., *Takeovers and Contests for Corporate Control*. New York: Oxford University Press, forthcoming. Also published in slightly abridged form in the *Midland Corporate Finance Journal*, Summer 1986.

Jensen, Michael C., and William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, 1976, *3*, 305–360.

Jensen, Michael C.,and R. Ruback, "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics*, April 1983, *11*, 5–50.

Jensen, Michael C., and Clifford Smith, Jr., "Stockholder, Manager, and Creditor Interests: Applications of Agency Theory." In Altman, Edward I., and Marti G. Subrahmanyam, eds. *Recent Advances in Corporate Finance*. Homewood, Illinois: Irwin, 1985, pp. 93–131.

Kaplan, Steven, "Management Buyouts: Efficiency Gains or Value Transfers," manuscript, Harvard Business School, November 1987.

Koleman, Joe, "The Proxy Pressure on Pension Fund Managers," *Institutional Investor*, July 1985, 145–147.

Lambert, R., and D. Larcker, "Golden Parachutes, Executive Decision-Making, and Shareholder Wealth," *Journal of Accounting and Economics*, April 1985, *7*, 179–204.

Langtieg, T. C., "An Application of a Three-Factor Performance Index to Measure Stockholder Gains from Merger," *Journal of Financial Economics*, December 1978, *6*, 365–384.

Magenheim, Ellen B., and Dennis Mueller, "On Measuring the Effect of Acquisitions on Acquiring Firm Shareholders or Are Acquiring Firm Shareholders Better Off After an Acquisition Than They Were Before?" Presented at the Conference on Takeovers and Contests for Corporate Control, Columbia University, November, 1985.

Malatesta, Paul H., and Ralph A. Walkling, "The Impact of Poison Pill Securities on Stockholder Wealth," *Journal of Financial Economics*, forthcoming.

Mandelker, Gershon, "Risk and Return": The Case of Merging Firms," *Journal of Financial Economics*, December 1974, *1*, 303–336.

Mergers and Acquisitions, May-June 1987, *21*, 16.

Merton, Robert C., "On the Current State of the Stock Market Rationality Hypothesis," Working Paper No. 1717-85, Sloan School of Management, MIT, October 1985.

Mikkelson, Wayne H., and Richard S. Ruback, "An Empirical Analysis of the Interfirm Equity Investment Process," *Journal of Financial Economics*, December 1985, *14*, 523–553.

Mikkelson, Wayne H., and Richard S. Ruback, "Targeted Repurchases and Common Stock Returns," manuscript, June 1986.

Morrison, Ann, "Those Executive Bailout Deals," *Fortune*, December 13, 1982, 82–87,

Murphy, Kevin J., "Corporate Performance and Managerial Remuneration: An Empirical Analysis," *Journal of Accounting and Economics*, April 1985, *7*, 11–42.

Palepu, Krishna G., "Predicting Takeover Targets: A Methodological and Empirical Analysis," *Journal of Accounting and Economics*, 1986, *8*, 3–35.

"R&D Scoreboard: Now, R&D is Corporate America's Answer to Japan Inc.," *Business Week*, June 23, 1986.

"R&D Scoreboard: Reagan & Foreign Rivalry Light a Fire Under Spending," *Business Week*, July 8, 1985, 86–87.

"R&D Scoreboard: Research Spending is Building Up to a Letdown," *Business Week*, June 23, 1987, 139–140.

Roll, Richard, "Empirical Evidence on Takeover Activity and Shareholder Wealth." In Coffee, John, Louis Lowenstein, and Susan Rose-Ackerman, eds., *Takeovers and Contests for Corporate Control*, New York: Oxford University Press, forthcoming.

Rozeff, M., "Growth, Beta and Agency Costs as Determinants of Dividend Payout Ratios," *Journal of Financial Research*, 1982, *5*, 249–59.

Ryngaert, Michael, "The Effect of Poison Pill Securities on Shareholder Wealth," *Journal of Financial Economics*, forthcoming.

Schipper, Katherine, and Abbie Smith, "Corporate Income Tax Effects of Management Buyouts," unpublished manuscript, University of Chicago, July 1986.

Shiller, Robert J., "Do Stock Prices Move Too Much to be Justified By Subsequent Changes in Dividends?" *American Economic Review*, 1981, 421–36.

Smith, Clifford W., Jr., and Jerold B. Warner, "On Financial Contracting: An Analysis of Bond Covenants," *Journal of Financial Economics*, 1979, *7*, 117–161.

Taggart, Robert A., "The Growth of the 'Junk' Bond Market and Its Role in Financing Takeovers," manuscript, Boston University, September 1986.

Wansley, James W., and A. Fayez, "Determinants of Return to Security Holders from Mergers," manuscript, Louisiana State University, 1986.

Wansley, James W., William R. Lane, and Ho C. Yang, "Abnormal Returns to Acquired Firms by Type of Acquisition and Method of Payment," *Financial Management*, Autumn 1983, *12*, No. 3, 16–22.

Wansley, James W., William R. Lane, and Ho C. Yang, "Gains to Acquiring Firms in Cash and Securities Transactions," *Financial Review*, forthcoming.

Weston, J. Fred., and Thomas E. Copeland, *Managerial Finance*. Chicago: Dryden Press, 1986, p. 649.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 49–68

The Market for Corporate Control: The Empirical Evidence Since 1980

Gregg A. Jarrell, James A. Brickley, and
Jeffry M. Netter

Corporate takeovers have been very big business in the 1980s. The Office of the Chief Economist (OCE) of the Securities and Exchange Commission estimates that shareholders of target firms in successful tender offers from 1981 through 1986 received payments in excess of $54 billion over the value of their holdings before the tender offers. Almost $38 billion of the total was received after 1984. If we include the increased wealth of target firm shareholders resulting from leveraged buyouts, mergers, and corporate restructurings (prompted in large part by the threat of takeovers) these numbers are even larger. W. T. Grimm & Co. collects similar data for a larger sample of change-of-control transactions, including mergers and leveraged buyouts. They estimate that from 1981 to 1986 the total dollar value of the premiums over the pre-announcement price paid for securities involved in change-of-control transactions was $118.4 billion.[1] Corporate restructurings have created even more

---

[1] These estimates understate the total premiums (dollar value of the percentage increase in the target's stock price caused by the takeover) paid in change-of-control transactions. OCE's sample is limited to the first successful tender offer for a firm. Thus, if an "auction" for the firm develops, resulting in an even higher offer price, they do not capture that additional premium. In addition, no account is made for tender offers that do not ultimately succeed. Shareholders may sell their shares in the market at the premium induced by the offer before it is known that the offer fails. The W. T. Grimm data also understates the total profits earned by stockholders because they calculate the premium based on the market price only five days before the initial public announcement and do not capture the premium attributable to increases in share prices that occur more than five days in advance of a public announcement.

■ *Gregg A. Jarrell is Senior Vice-President, Alcar Group Inc., Skokie, Illinois; James A. Brickley is Associate Professor, William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York; Jeffry M. Netter is Senior Financial Economist, Office of the Chief Economist; United States Securities and Exchange Commission, Washington, DC.*

wealth. For example, Jensen (1986) estimated that the restructurings of Phillips, Unocal and Arco created total gains to shareholders of $6.6 billion by reducing investment in negative net present value projects.

There are numerous factors behind the high level of takeover activity in the 1980s. For example, antitrust regulators have come to understand that in the increasingly competitive international marketplace U.S. interests are well-served by domestic mergers that could be objectionable in a more closed economy. Today's antitrust regulators almost never object to vertical combinations, and even horizontal mergers between industry leaders—completely taboo before the 1980s—are often allowed today.

Deregulation also has induced merger and acquisition activity by calling forth new skills and strategies, and new management teams to implement them. Many of the mergers, takeovers, and restructurings over the last ten years have occurred in industries that recently were deregulated such as airlines and transportation, financial services, broadcasting, and oil and gas. For example, transportation and broadcasting together accounted for 20 percent of all mergers and acquisition activity from 1981 and 1984 while oil and gas accounted for another 26.3 percent (Jensen, 1986).

Other factors motivating the high level of takeover and restructuring activity in the 1980s include innovations in takeover financing, less potent state antitakeover regulations, the retreat by the Federal courts and regulatory agencies from protecting besieged target firms, and learning about the possible returns to this type of activity. These factors are critical to understanding why firms that were considered "untouchable" not long ago have been the targets of hostile takeovers with increasing frequency. This growing list includes USX, CBS, Phillips, and TWA, to name just a few.

The Council of Economic Advisors (CEA) in the 1985 *Economic Report of the President* provides data on the extent of the takeover activity in the 1980s and the importance of large transactions in explaining this activity. The CEA states that the increase in merger and acquisition activity in the 1980s is due to a large increase in the size of the largest transactions. Their evidence indicates that in the period 1981 to 1984 the average annual reported real value of mergers and acquisitions was 48 percent greater than in any four year period from the late 1960s to the early 1970s. In addition, of the 100 largest acquisition transactions recorded through 1983, 65 occurred after 1982 and only 11 took place prior to 1979.

Returns to Bidders and Targets

Critics of takeovers question whether tender offers, mergers, and leveraged buyouts produce net gains to society. Critics argue any gains to a given party are simply redistributions resulting from losses to someone else (or more colorfully put, a pirating of assets by modern financial buccaneers). Also critics contend that battles for corporate control divert energy from more productive endeavors.[2] In this section, we find that such criticisms are ill founded, and thus conclude that battles for corporate control serve a beneficial function for the economy.

The market for corporate control is the market for the right to control the management of corporate resources. In a takeover, an outside party seeks to obtain control of a firm. There are several types of takeovers, including mergers, hostile and friendly tender offers, and proxy contests. In a merger the bidder negotiates an agreement with target management on the terms of the offer for the target and then submits the proposed agreement to a vote of the shareholders. In a tender offer, a bidder makes an offer directly to shareholders to buy some or all of the stock of the target firm. A "friendly" tender offer refers to offers that are supported by target management. The most controversial type of takeovers are "hostile" tender offers, which are tender offers that are opposed by target managements. In a proxy contest, a dissident group attempts through a vote of shareholders to obtain control of the board of directors. Finally, leveraged buyouts are buyouts of shareholder's equity, heavily financed with debt by a group that frequently includes incumbent management.

Many of the studies reviewed in this paper are event studies that measure the effects of certain unanticipated events (such as a takeover or other control contest) on stock prices, after correcting for overall market influence on security returns. Any finding of abnormal returns, therefore, shows how the stock market views the impact of the event on the firm's common stockholders. (See Brown and Warner, 1985, for a more thorough review of event study methods.)

Returns to Shareholders of Target Companies

Shareholders of target companies clearly benefit from takeovers. Jarrell and Poulsen (1987a) estimate the premiums paid in 663 successful tender offers from 1962 to December 1985. They find that premiums averaged 19 percent in the 1960s, 35 percent in the 1970s, and from 1980 to 1985 the average premium was 30 percent. These figures are consistent with the 13 studies of pre-1980 data contained in Jensen and Ruback (1983) which agree that targets of successful tender offers and mergers before 1980 earned positive returns ranging from 16 percent to 30 percent for tender offers.[3]

Similar results are contained in studies of leveraged buyouts and going private transactions. Lehn and Poulsen (1987) find premiums of 21 percent to shareholders in

---

[2] Many critics of acquisition activity (such as the Business Roundtable) are primarily concerned with alleged abuses arising from hostile takeovers. The other types of acquisition activity are approved by target firms' management who allegedly are the individuals most concerned with the welfare of the target firm and its shareholders.

[3] Jensen and Ruback (1983) review 13 studies published between 1977 and 1983—six on mergers and seven on tender offers. Their survey provides a concise summary of the pre-1980 data. But because of the lengthy review process for academic journals, the most up-to-date sample used in these studies ends in 1981, and most do not go beyond the late 1970s. This paper can be considered an update of Jensen and Ruback with a focus on recent empirical studies that cover takeovers made in the 1980s.

93 leveraged buyouts taking place from 1980 to 1984. DeAngelo, DeAngelo and Rice (1984) find an average 27 percent gain for leveraged buyouts between 1973 and 1980.

OCE (1985a) measures premiums paid by comparing the price per share offered by the bidder to the trading price of the stock one month before the offer, not adjusting for changes in the market index (also see Comment and Jarrell, forthcoming). Using a comprehensive sample of 225 successful tender offers from 1981 through 1984, including over-the-counter targets, OCE finds the average premium to shareholders to be 53.2 percent. OCE has updated these figures for 1985 and 1986 and finds a decrease over the last two years. OCE finds that the average premium is 37 percent in 1985 and 33.6 percent in 1986.[4]

While the evidence reported thus far indicates substantial gains to target shareholders, it probably understates the total gains to these shareholders. In many cases events occur before a formal takeover offer, so studies that concentrate on the stock price reactions to formal offers will understate the total gains to shareholders.

Several recent empirical studies examine the stock market reaction to events that often precede formal steps in the battle for corporate control. Mikkelson and Ruback (1985a) provide information on the stock price reaction to Schedule 13D filings. Schedule 13D must be filed with the SEC by all purchasers of 5 percent of a corporation's common stock, requiring disclosure of, among other things, the investor's identity and intent. Mikkelson and Ruback find significant price reactions around the initial announcement of the filing, and that the returns depend on the intent stated in the 13D. The highest returns, an increase of 7.74 percent, occurred when the filer in the statement of intent indicated some possibility of a control change. However, the abnormal returns were only 3.24 percent if the investor reported the purchase was for investment purposes. Holderness and Sheehan (1985) find a differential stock market effect to 13D filings depending on the identity of the filer. They show the filings of six "corporate raiders" increased target share prices by a significantly greater amount than a sample of other filers (5.9 percent to 3.4 percent).

More direct evidence that significant stock price increases occur prior to formal announcements of corporate events is contained in OCE (1987c) which finds a significant increase in the stock price of target firms in 172 successful tender offers in the period before any announcement of the offer. OCE finds a run-up in stock prices of 38.8 percent of the total control premium by the close of day *before* the offer announcement. The announcement date is, in the parlance of Wall Street, the date the target firm was put "in play" and represents some event having significant implications for corporate control. For example, the in-play date in some cases is the ormal offer but in other cases is the eventual bidder's filing of a Schedule 13D with orporate control implications for the target.

The OCE (1985a) study also explicitly tests and rejects the popular theory that two-tier tender offers isadvantage target shareholders. Some observers argue that two-tier offers—in which the bidder first akes an offer for control of the firm and then makes a "clean-up" offer for remaining shares at a lower rice—coerces shareholders to tender to avoid the clean-up price. OCE finds that two-tier offers have erall premiums that are nearly identical to the average for any-or-all offers and that there is no evidence at two-tier offers "stampede" shareholders into unwise trading decisions.

Table 1
Cumulative excess returns to successful bidders for tender offers during 1960 to 1985, by decade

Trading-day Interval	*Cumulative excess returns in percent* *All*	*1960s*	*1970s*	*1980s*
− 10 to +5	1.14	4.40	1.22	−1.10
(*t*-stat.)	(2.49)	(4.02)	(2.12)	(−1.54)
−10 to +20	2.04	4.95	2.21	−0.04
(*t*-stat.)	(3.31)	(3.52)	(2.87)	(−0.04)
Number of observations	405	106	140	159

Source: Jarrell and Poulsen (1987a)

While some commentators argue that price run-up before the formal announcements of tender offers indicates the presence of illegal insider trading, OCE's evidence demonstrates that the legal market for information can explain much of the run-up. OCE shows that a significant portion of the run-up can be explained by three readily identifiable influences on pre-bid trading: media speculation, the bidder's foothold acquisition in the target, and whether the bid is friendly or hostile. Systematic relations between these factors and run-up in target share prices indicate that there is an active market for information about impending takeover bids and a large portion of the run-up can be explained by factors other than illegal insider trading. OCE's results on pre-bid market activity are supported by Comment (1986).

Returns to Shareholders of Acquiring Companies

The 1980s evidence on bidders comes from Jarrell and Poulsen (1987a), with data on 663 successful tender offers covering 1962 to 1985. Table 1 summarizes the excess returns to 440 NYSE and AMEX bidders. For the entire sample period bidders on average realized small, but statistically significant, gains of about 1 to 2 percent in the immediate period around the public announcement. Most interesting is the apparent secular decline in the gains to successful bidders in tender offers. Consistent with the previous studies reviewed by Jensen and Ruback (1983), Table 1 shows positive excess returns of five percent during the 1960s, and a lower, but still significantly significant, positive average of 2.2 percent over the 1970s. However, the 159 cases from the 1980s show statistically insignificant losses to bidders.

How the Distribution of Takeover Gains Is Determined

Companies that are targets of takeovers receive the bulk of the value created by corporate combinations and these gains are not offset by losses to acquirers. As one might predict, an important factor in determining how these takeover gains are split

This evidence indicates that the market does not, on average, learn much of anything that is new or different about target firms' intrinsic values through the tender offer process, despite the tremendous attention lavished on targets, and the huge amounts of information traded among market participants during takeover contests. If undervaluation had indeed been present, then the deluge of new information on the intrinsic value of targets should have caused fundamental price corrections even in the event of takeover defeats. But in the overwhelming majority of cases studied, prices dropped rather than increased for target firms that fought off takeovers.

Do Tax Effects Motivate Mergers and Takeovers? Tax motives have long been suspected as an important cause of merger and acquisition activity. Indeed, the Tax Reform Act of 1986 contains several provisions aimed at reducing the tax benefits available through mergers.[9] Most recent studies, however, assign tax benefits a minor role in explaining merger and takeover activity. Auerbach and Reishus (1987a) study 318 mergers and acquisitions during 1968–83 to estimate the tax benefits available in these transactions from increased use of tax losses and credits. They found that these tax benefits in general were not a significant factor in the majority of large acquisitions. In a fair number of transactions, however (potentially 20 percent of the mergers), tax factors did appear to be significant enough to affect the decision to merge.[10] Lehn and Poulsen (1987) find, in their study of leveraged buyouts from 1980–84, that the premiums paid are directly related to potential tax benefits associated with these transactions, suggesting that in part these leveraged buyouts are motivated by tax considerations.

In summary, acquiring firms' tax losses and credits, and the option to step-up the basis of targets' assets without paying corporate level capital gains, are two tax benefits that appear to have had some impact on merger activity. However, the evidence suggests that much of the takeover activity in the last twenty years was not tax motivated.

Do Bondholders Lose From Takeovers? Some critics of takeovers suggest that the premiums paid by bidders are not a result of any wealth enhancing changes, but instead represent a redistribution from the holders of the target's bonds and preferred equity. For example, the bonds of an acquiring firm can drop in value if the acquiring firm pays cash for a riskier target firm. Given that the combined value of the two firms remains unchanged, the decline in the bond value will be captured as a gain by some other class of security holder (such as common stockholders).[11] However, the empirical evidence does not support this argument.

---

[9] A review of the effects on the takeover market of the change of the 1986 Tax Reform Act is contained in Steindel (1986). The 1986 Tax Reform Act repeals the General Utilities doctrine, which states that corporations liquidating their businesses are not subject to capital gains tax on the value of their assets. A firm using General Utilities in a liquidation (the purchaser of at least 80 percent of the stock of a corporation may treat the transaction for tax purposes as a liquidation) avoids the tax liability that comes with appreciated assets. Steindel argues that the repeal of the General Utilities doctrine combined with the changes in corporate tax rates reduces the attractiveness of many mergers and acquisitions. Other tax changes with effects on takeover activity include the increase in the personal capital gains tax and new rules on the transfer of net operating loss carryforwards.

[10] However, Auerbach and Reishus (1987b) compare actual mergers over 1968–1983 with a control group of nonmerging firms and conclude that the potential increase in interest deductions and unused tax losses and tax credits of the acquired firms have not driven acquisitions.

Denis and McConnell (1986) examine the returns on various classes of the securities of a sample of 132 mergers in the period 1962 to 1980. Denis and McConnell's results are consistent with earlier studies in that they find gains to mergers and no losses to bondholders. Their results indicate that on average holders of common stock, convertible and nonconvertible preferred stock, and convertible bonds in the acquired firm gain from a merger. Those who hold nonconvertible bonds in the acquired firm and convertible bonds, nonconvertible bonds, and nonconvertible preferred stock in the acquiring firm neither gain nor lose in a merger. Denis and McConnell also find some evidence that the acquiring firms' common shareholders do not lose and may gain from mergers, especially in the days immediately following the announcement. Lehn and Poulsen's (1987) study of 108 leveraged buyouts from 1980 to 1984 finds no support for the redistribution theory. They find no evidence that the shareholder value created by the leveraged buyouts comes at the expense of preferred shareholders or bondholders. In sum, the evidence provides no support for the hypothesis that the supposed gains from acquisitions are actually transfers from the holders of senior securities to the holders of common stock.

Do Labor's Losses Finance Takeovers? Recent takeovers in the airline industry have involved conflict between acquiring-firm management and the (usually) unionized labor of the target firm. These conflicts have contributed to the popular generalization that shareholder premiums from takeovers come largely at the expense of labor. Shleifer and Summers (1987) articulate this view more rigorously focusing on implicit long-term contracts between labor and incumbent (target) management. They argue raiders can sometimes exploit these contracts by buying a controlling share of the equity and financing the premium by using pressure tactics to force significant wage concessions. In theory, this activity can be socially inefficient by ruining the market for these implicit long-run labor contracts and forcing labor and management to use less efficient contracting devices.

This redistributive theory from labor to shareholders has not been tested widely, but a recent NBER study by Brown and Medoff (1987) presents statistical evidence based on Michigan's employment and wages that fails to support it. Although this close look at Michigan is not necessarily indicative of the U.S. experience (for example, it contains few large mergers or hostile tender offers), the results are that wages and employment rise on average for firms that are involved in acquisitions.

Summary of Source of Gains

The various redistribution theories of takeover gains have been the subject of considerable empirical work since the Jensen and Ruback (1983) review. Most convincing is the empirical rejection of the undervaluation theory: target firms cannot be depicted generally as being "undervalued" by the stock market. Also soundly rejected by the data is the short-term myopia theory. The evidence gives tax-benefits theories at least a minor role in explaining merger and tender offer activity. Finally, evidence is inconsistent with the theories that the stock-price gains to shareholders come from bondholders and labor.

---

[11] The senior security holders of the target firm can also lose depending on the takeover's effect on the riskiness of their claims.

Although some individuals (incumbent management, for example) obviously lose in at least some takeovers, the literature, while not conclusive, offers little or no support for the notion that the redistribution theories explain a major portion of the apparent gains from takeovers. It has been impossible so far to find systematic losses which could offset the enormous gains to target and bidding firm shareholders from mergers, tender offers, and other corporate-control activities. We therefore conclude that evidence is consistent with the notion that these corporate transactions reflect economically beneficial reshufflings of productive assets.

The Effects of Defending Against Hostile Takeovers

Defensive strategies against hostile takeovers have always been controversial since they pose a conflict of interest for target management. After all, takeovers can impose significant welfare losses on managers, who may be displaced and lose their organization-specific human capital. These conflicts may tempt some managers to erect barriers to hostile takeovers, thus insulating themselves from the discipline of the outside market for control at the expense of their shareholders and the efficiency of the economy.

However, providing target management with the power to defend against hostile takeover bids might also help target shareholders during a control contest. Target management can in certain cases defeat bids that are "inadequate." Although this rationale is popular, the evidence discussed earlier shows that in very few cases do these alleged long-term gains of independence actually materialize. The other benefit of resistance comes when resistance by target management helps promote a takeover auction. Litigation and other blocking actions can provide the necessary time for the management of the target firm to "shop" the target and generate competing bids. This auction rationale for resistance is harder to reject statistically. Evidence on occasional shareholder losses after the defeat of a takeover attempt does not in itself disprove the auction theory. This negotiating leverage can be expected to fail in some cases, with the sole bidder becoming discouraged and withdrawing. It is a gamble. The hypothesis is rejected only if the harmful outcome of defeating all bids is sufficiently frequent and costly to offset the benefits of inducing higher takeover prices. One must also consider the social cost of tender offers that never occur because of the presence of defensive devices. Unfortunately, this deterrence effect is very difficult to measure and we present no direct evidence of the extent of these costs.

Evidence on the effects of defensive measures by target management is obtained mainly from two approaches, the event-type study and the outcomes-type study. The event-type study recognizes that an efficient market must judge this cost-benefit tradeoff when it adjusts the market value of a firm in response to the adoption of a charter amendment or some other kind of resistance. Alternatively, the outcomes-type study examines the actual outcomes of control contests over a significant time horizon among firms using a common kind of resistance—say all firms adopting poison pills. That is, an event study measures the stock price reaction to the introduction of defensive devices while outcomes studies follow the use of defensive devices in control contests to determine their effects on the outcomes of the contests.

Many defensive measures must be approved by a vote of the shareholders. Hence, voting has the potential to block management-sponsored proposals that harm shareholders, depending on the costs and benefits to individual shareholders from collecting relevant information and voting. In general, a shareholder with a small amount of shares will not invest heavily in the voting process since a small number of shares will not generally affect the outcome regardless of how they are voted. However, if individual voting and information costs are near zero even the shareholder with few shares can be expected to vote against management on value-decreasing proposals. Alternatively, large outside block holders (like institutional investors) internalize more of the benefits from participation in the voting process and can be expected to take an active interest in voting on antitakeover proposals even when the information gathering and voting costs are positive. Since voting rights can block harmful measures, we distinguish between two broad categories of defensive measures, those receiving approval by voting shareholders and those adopted unilaterally by management.

Defensive Measures Approved by Shareholders

Antitakeover amendments generally operate by imposing new conditions that must be satisfied before changing managerial control of the corporation. They are almost always proposed by management and they usually require majority voting approval by shareholders. Proposed antitakeover amendments are very rarely rejected by voting shareholders; Brickley, Lease, and Smith (forthcoming) find for a sample of 288 management-sponsored antitakeover proposals in 1984 that about 96 percent passed.

Supermajority Amendments. Most state corporation laws set the minimum approval required for mergers and other important control transactions at either one-half or two-thirds of the voting shares. Supermajority amendments require the approval by holders of at least two-thirds and sometimes as much as nine-tenths of the voting power of the outstanding common stock. These provisions can apply either to mergers and other business combinations or to changing the firm's board of directors or to both. Pure supermajority provisions are very rare today, having been replaced by similar provisions that are triggered at the discretion of the board of directors. This allows the board to waive the supermajority provisions allowing friendly mergers to proceed unimpeded.

Five years ago, Jensen and Ruback (1983) found mixed evidence on the effect of supermajority amendments passed before 1980. However, a more recent study by Jarrell and Poulsen (1987b), derived from OCE (1985c), covers 104 supermajority amendments passed since 1980 and reports significant negative stock-price effects of over 3 percent around the introduction of the proposals. They also show that firms passing supermajority amendments have relatively low institutional stockholdings (averaging 19 percent) and high insider holdings (averaging 18 percent), which they interpret as helping to explain how these amendments received voting approval despite their harmful wealth effect. That is, firms proposing these amendments have fewer blockholders with incentives to invest in the voting process. Jarrell and Poulsen further conjecture that the increased shareholder resistance to harmful supermajority amendments helps explain their declining popularity in contrast to the success of the fair price amendment which appear less likely to harm shareholders (as discussed below).

Fair Price Amendments. The fair price amendment is a supermajority provision that applies only to nonuniform, two-tier takeover bids that are opposed by the target's board of directors. Uniform offers that are considered "fair" circumvent the supermajority requirement, even if target management opposes them. Fairness of the offer is determined in several ways. The most common fair price is defined as the highest price paid by the bidder for any of the shares it has acquired in the target firm during a specified period of time. Jarrell and Poulsen (1987b) report that 487 firms adopted fair price charter provisions between 1979 and May 1985, with over 90 percent of these coming in the very recent period of 1983 to May 1985.

The stock price effects reflect the low deterrence value of the fair price amendment. Jarrell and Poulsen (1987b) report an average loss of 0.73 percent around the introduction of these amendments, which is not statistically significant. They also show that firms adopting fair price amendments have roughly normal levels of insider holdings (12 percent) and of institutional holdings (30 percent). They interpret this evidence as supporting the view that shareholder voting retards adoption of harmful amendments, especially when insider holdings are low and institutional holdings are high. Further support for this view is provided by Brickley, Lease, and Smith (forthcoming) who document that "no" votes on antitakeover amendments (especially ones that harm shareholders) increase with institutional and other outside blockholdings, while "no" votes decrease with increases in managerial holdings.

Dual-Class Recapitalizations. These plans restructure the equity of a firm into two classes with different voting rights. Although several methods are used, the common goal is to provide management or family owners with voting power disproportionately greater than provided by their equity holdings under a "one share-one vote" rule.[12]

Evidence before and after 1980 has confirmed that the market generally values shares with voting power more than those without. Lease, McConnell, and Mikkelson (1983) examine 30 firms having dual-class common stock and show that voting stock on average trades at a significant premium, ranging from one to seven percent. A recent paper by OCE (1987a) examines the monthly stock prices of 26 OTC and

---

[12] For over 60 years, the New York Stock Exchange did not allow any member firm to have a dual-class capitalization structure, but it has recently proposed a liberalization to allow dual-class listings in response to competitive pressures from Amex and OTC markets. Amex currently allows dual-class listings with some restrictions, and the OTC market has no restrictions beyond usual state-law requirements.

AMEX firms having dual-class common and shows an average discount of four to five percent for low-vote common, though the discount is reduced when the low-vote stock has rights to preferential dividends.

Of course, the fact that the market values voting power does not demonstrate that dual-class recapitalizations reduce the overall price of stock. DeAngelo and DeAngelo (1985) examine in detail 45 firms that had dual-voting common stock as of 1980. They find that, after restructuring, management and family insiders control a median of 57 percent of the votes and 24 percent of the common stock cash flows. This confirms that dual-class structures often confer substantial voting powers on incumbent management. However, DeAngelo and DeAngelo also suggest that the shareholders of the firms in this sample found it beneficial to contract with incumbent management to limit the competition for management of their firms. They argue that shareholders rationally accept a reduced potential for hostile takeovers in return for other benefits, such as greater incentives for incumbents to make specific long-term investments in human capital.

Two recent studies have addressed the empirical question of whether dual-voting structures are beneficial, as DeAngelo and DeAngelo suggest, or harmful to outside shareholders. Partch (1987) examines the stock-price reaction around the announcement of the proposed dual-class recapitalizations for 44 firms. She reports nonnegative share-price effects. However, for more recent recapitalizations, Jarrell and Poulsen (forthcoming, extending OCE 1987a, 1987d) find negative effects at the announcement of dual-class recapitalizations. For a sample of 89 firms delisting from 1976 through 1987, they report an average abnormal stock price effect of −.93 percent.

If dual-class recapitalization proposals are viewed primarily as takeover defenses, their announcement should cause negative stock price reactions, similar to those observed at the announcement of supermajority amendments. However, firms announcing dual-class recapitalizations have some unusual characteristics. Jarrell and Poulsen (forthcoming) find that the average net-of-market return to their 94 dual-class firms over the year preceding the recapitalization is over 37 percent. Jarrell and Poulsen and Partch both find that insider holdings average 44 percent before the recapitalization, and that recapitalization significantly increases insider voting control. These two characteristics suggest that the typical dual-class firm is already controlled by insiders and the recapitalization provides a means to raise needed capital for positive net present value projects without the dilution of control.

Changes in the State of Incorporation. Changing the state of incorporation can affect the contractual arrangements between management and shareholders. For example, some states such as Ohio, Indiana, and New York have elements in their corporate codes that make takeovers more difficult than in other states. Dodd and Leftwich (1980) find that firms change their state of incorporation after a period of superior performance and that the change itself is associated with small positive excess returns. More recently, Romano (1985) finds a statistically significant price increase around the reincorporation announcement in a sample of firms that reincorporate for various reasons. However, in the subsample of 43 firms who reincorporated as an antitakeover device she found a small statistically insignificant price increase at the announcement of a reincorporation. The evidence is not conclusive but it does indicate that reincorporating in a new state does not on average harm shareholders.

Reduction in Cumulative Voting Rights. Cumulative voting makes it possible for a group of minority shareholders to elect directors even if the majority of shareholders oppose their election. Dissidents in hostile takeovers and proxy contests will often attempt to elect some board members through the use of cumulative voting. Bhagat and Brickley (1984) examine the stock price reaction to 84 management sponsored charter amendments that either eliminate or reduce the effect of cumulative voting. Since these amendments decrease the power of dissident shareholders to elect directors, they increase management's ability to resist a tender offer. Bhagat and Brickley find statistically significant negative abnormal returns of about one percent at the introduction of these charter amendments.

Defensive Measures That Do Not Require Shareholder Approval

Four general kinds of defensive measures do not require voting approval by shareholders: general litigation, greenmail, poison pills, and the use of state antitakeover laws. With the exception of general litigation, these defensive actions are associated on average with negative stock-price reactions indicating that in most cases they are economically harmful to stockholders of companies whose management enacted them.

Litigation by Target Management. As described earlier, litigation can be expected to hurt shareholders of some target companies by eliminating takeovers and to help shareholders of other companies by giving their management time and weapons to cut a better deal. Jarrell (1985) examines 89 cases involving litigation against a hostile suitor based on charges of securities fraud, antitrust violations, and violations of state or Federal tender offer regulations. His results show that litigation usually delays the control contest significantly and that litigating targets are frequently the beneficiaries of auctions. The 59 auction-style takeovers produced an additional 17 percent excess return to shareholders over the original bid, while the 21 targets that remained independent lost nearly all of the original average premium of 30 percent. Overall, Jarrell concludes that this evidence cannot reject the theory that on average target litigation is consistent with shareholder wealth maximization.[13]

However, harm can result from certain types of defensive litigation. Netter (1987) finds that litigation based in part on a claim alleging the filing of a false Schedule 13D Item 4 can be detrimental to target shareholders. In an exhaustive sample of all cases where target management filed a suit alleging (among other things) that a bidder filed a false 13D Item 4, he finds that target shareholders are better off if their management loses the case than if they win. If the target firm wins the case its share price declines by a significant amount (an abnormal return of negative 3.37 percent in the two-day window around the decision) while if the bidder wins, the stock price of the target firm increases by a significant amount (positive 3.15 percent abnormal return in the two-day window.

---

[13] Jarrell also notes that while defensive litigation redistributes premiums it also, by reducing incentives to engage in takeovers and through the cost of the litigation itself, can reduce social welfare.

Targeted Block Stock Repurchases (Greenmail). Greenmail occurs when target management ends a hostile takeover threat by repurchasing at a premium the hostile suitor's block of target stock. This controversial practice has been challenged in federal courts, in congressional testimony, and in SEC hearings, and it has brought negative publicity both to payers and to receivers of greenmail. In reviewing earlier studies, Jensen and Ruback (1983) conclude that greenmail repurchases are associated with significantly negative abnormal stock returns for the shareholders of the repurchasing firms (probably because they eliminate potential takeover bids) and significantly positive abnormal stock returns for shareholders of the selling firms. These negative effects of greenmail repurchases contrast sharply with the normally positive stock-price effects associated with nontargeted offers to repurchase a company's own stock.

Since then, three new empirical studies have contributed to a more complex and less conclusive discussion of greenmail transactions. These studies indicate that it is not necessarily in the interests of shareholders to ban greenmail payments. Such a ban has the potential to discourage outside investment in the potential target's stock by investors anticipating greenmail payments and hence reduces the incentives of outsiders to monitor managers.

Mikkelson and Ruback (1985a) examine 39 cases of greenmail (based on 13Ds filed during 1978–80). They find a significant stock-price loss of 2.3 percent upon the announcement of the repurchases. However, they also report an average gain of 1.7 percent over the entire period including the original stock purchase by the hostile suitors. Holderness and Sheehan's (1985) outcome-type study includes 12 cases of greenmail, and they report a pattern of returns consistent with the evidence of Mikkelson and Ruback. Although the greenmail transaction itself harms target shareholders, the net returns to stockholders resulting from the initial purchase and related events is positive. A more comprehensive sample of targeted block stock repurchases is covered by OCE (1984). This study includes 89 cases of large repurchases (blocks greater than 3 percent of the outstanding common stock) from 1979 to 1983. The initial announcement of investor interest induces a positive return averaging 9.7 percent, while the greenmail transaction is associated with a stock price loss of 5.2 percent.

Poison Pills. Since its introduction in late 1982, the "poison pill" has become the most popular and controversial device used to defend against hostile takeover attempts. Poison pill describes a family of shareholder rights agreements that, when triggered by an event such as a tender offer for control or the accumulation of a specified percentage of target shares by an acquirer, provide target shareholders with rights to purchase additional shares or to sell shares to the target at very attractive prices. These rights, when triggered, impose significant economic penalties on a hostile acquirer.

Poison pills are considered very effective deterrents against hostile takeover attempts because of two striking features. First, pills can be cheaply and quickly altered by target management if a hostile acquirer has not pulled the trigger. This feature pressures potential acquirers to negotiate directly with the target's board. Second, if not redeemed, the pill makes hostile acquisitions exorbitantly expensive in most cases. As an obstacle to hostile takeover attempts, the poison pill is unmatched except by dual-voting recapitalizations or direct majority share ownership by incumbent management. The concern over poison pills was heightened by the Delaware Supreme Court's 1985 ruling in *Moran v. Household International*[14] that poison pills do not require majority voting approval by shareholders.

The most comprehensive study of poison pills is Ryngaert (forthcoming), which is an outgrowth of OCE (1986). The Ryngaert study features an exhaustive collection of 380 poison pills adopted from 1982 to December 25, 1986. Over 80 percent of these were adopted after the *Household* decision. Ryngaert divides his sample into discriminatory pills (the most restrictive) and flip-over pills (the least restrictive). He also accounts for whether firms are subject to takeover speculation and whether confounding events occur close to the announcement of the pill that contaminate the data. The stock-price effect over the 283 cases with no confounding events is a statistically significant −.34 percent. Focusing on 57 cases subject to takeover speculation, the average loss is 1.51 percent, also statistically significant. These results are supported by the findings of Malatesta and Walkling (forthcoming).

Discriminatory pills have more harmful effects on shareholder wealth than do flip-over pills. Also, the discriminatory pills that threaten the hostile suitor with severe dilution have become increasingly popular. Ryngaert reports that pill-adopting managements own a surprisingly low average of around 3.0 percent of their firms' outstanding stock. This fact, together with high institutional holdings, suggest that many of these firms would have difficulty obtaining shareholder voting approval if it were required.

Ryngaert also examines the stock-price effects of important court decisions emanating from legal battles involving pill defenses during 1983–86. He shows that 15 of 18 pro-target, pro-poison pill decisions have negative effects on the target's stock price, and 6 of 11 pro-acquirer decisions have positive effects on the target stock price. This evidence is inconsistent with the theory that pill defenses improve shareholder wealth by strengthening management's bargaining position in control contests.

Although these losses are not large in percentage terms, these empirical tests suggest that poison pills are harmful to target shareholders.

State Antitakeover Amendments. In addition to the Williams Act at the Federal level, tender offers are regulated by many states. So-called first-generation state antitakeover regulations are antitakeover laws that were passed by the states before the 1982 Supreme Court decision in *Edgar v. Mite*.[15] The Jarrell and Bradley (1980) study of state and Federal regulation of tender offers finds that first generation state regulations significantly increase the premiums paid in tender offers. Smiley (1981) illustrates the deterrent effects of these early state takeover regulations.

However, first generation antitakeover laws were generally extinguished in *Edgar v. Mite* when the Supreme Court ruled the Illinois antitakeover law unconstitutional. Justice White's opinion held the Illinois takeover statute was preempted by the Williams Act and constituted an undue and direct burden on interstate commerce. As a result over 20 states passed second generation antitakeover laws to attempt to pass constitutional muster under the Supreme Court's reasoning. While some of them have also been ruled unconstitutional, the Supreme Court in 1987 (*CTS v. Dynamics Corp. of America*) ruled the Indiana antitakeover law constitutional.[16] This decision already is leading states to pass third generation antitakeover laws that would be constitutional under the CTS reasoning.

---

[14] *Moran v. Household International*, 490 A.2d 1059 (1985).

[15] *Edgar v. Mite*, 457 U.S. 624, 102 S. Ct. 2629 (1982).

Two recent studies, Ryngaert and Netter (1987, based on OCE, 1987c) and Schumann (1987) provide more direct evidence on the wealth effects of state antitakeover regulations. Ryngaert and Netter examine the stock price effects of the passage of the Ohio antitakeover law on shareholders of firms chartered in Ohio. This act was passed during (and apparently motivated by) Sir James Goldsmith's attempted hostile takeover of Goodyear. They find that the passage of the law was accompanied by a significant stock-price loss of up to 3.24 percent to the shareholders of firms incorporated in Ohio with less than 30 percent inside ownership. This evidence on the impact of state takeover laws is supported by Schumann (1987) who finds a decline of approximately one percent to shareholders of New York firms on the announcement and passage of a New York antitakeover law. While these laws potentially could be beneficial to the individual states (if jobs are kept in the state by preventing takeovers), shareholders are harmed by state antitakeover regulations.

Summing Up Defensive Tactics

Four years ago Jensen and Ruback (1983) reviewed empirical studies of antitakeover charter amendments, shark repellents, changes of incorporation, and greenmail. They conclude (p. 47): "It is difficult to find managerial actions related to corporate control that harm stockholders; the exception are those actions that eliminate an actual or potential bidder, for example, through the use of targeted large block repurchases or standstill agreements."

Since their review, the defensive arsenal available to target management has been strengthened. These defensive tactics have been developed through a fascinating process of sequential innovations, as specific defenses arise to counter improved bidder finances and other tactics. In 1983, the now common fair-price amendment was a novel idea and the poison pill was not yet invented. Financial economists in academia and government have kept close pace with these developments, providing timely analyses of new charter amendments, poison pill defenses, greenmail transactions, and so on. While Jensen and Ruback were correct in predicting this area would be a "growth industry," we cannot reiterate their then-accurate conclusion that harmful defensive tactics are rare.

Conclusion

In the 1980s, the market for corporate control has been increasingly active, and the quantity of output of academic researchers studying corporate control questions has mirrored the market activity. This review has confirmed the basic conclusions of Jensen and Ruback's (1983) review article and has shed light on some questions Jensen and Ruback were forced to leave unanswered. Financial researchers continue to find larger premiums being paid to target shareholders for later tender offers than for earlier tender offers. Acquirers, however, receive at best modest increases in their stock price, and the winners of bidding contests suffer stock-price declines as often as they do gains. This pro-target division of takeover gains appears to be partially a result of improved defensive tactics that can effectively delay execution of the bid and allow the target to receive improved bids from others or fashion a defensive restructuring and stock buyback. The evidence further suggests that the premiums in takeovers represent real wealth gains and are not simply wealth redistributions.

---

[16] *CTS v. Dynamics Corp. of America*, 107 S. Ct. 1637 (1987).

Prominent in the 1980s are new studies of defensive measures, such as antitakeover charter amendments, targeted block stock repurchases (greenmail), dual-voting recapitalizations, state antitakeover laws, and poison pills. The general finding, although it is far from conclusive, is that defensive measures that require shareholder voting approval are less likely to be harmful to shareholder wealth than are defensive measures not subject to shareholder approval. Fair-price charter amendments and dual-class recapitalizations that require shareholder approval are not shown to be harmful to stock value, while poison pills and greenmail-type repurchases that do not need shareholder approval appear on average to reduce shareholder value. However, some proposals that require a favorable vote from shareholders to implement (e.g., supermajority provisions and the elimination of cumulative voting) on average appear to reduce shareholder wealth. These findings raise serious questions about whether the business judgment rule is operating too broadly as a shield for defensive actions by target managements.

■ *We are especially grateful to Joseph Stiglitz, Timothy Taylor, and Annette Poulsen for their many helpful comments and suggestions. The Securities and Exchange Commission as a matter of policy disclaims responsibility for any private publication or statement by any of its employees. The views expressed here are those of the authors and do not necessarily reflect the views of the Commission or the authors' colleagues on the Staff of the Commission.*

References

Auerbach, Alan J. and David Reishus, "Taxes and the Merger Decision." In Coffee, J. and Louis Lowenstein, eds., *Takeovers and Contests for Corporate Control*, Oxford: Oxford University Press, 1987a.

Auerbach, Alan J. and David Reishus, "The Effects of Taxation on the Merger Decision," NBER Working Paper, 1987b.

Bhagat, Sanjai, and James Brickley, "The Value of Minority Shareholder Voting Rights," *Journal of Law and Economics*, 1984, *27*, 339–365.

Bhagat, Sanjai, James Brickley, and Uri Lowenstein, "The Pricing Effects of Inter-Firm Cash Tender Offers, *Journal of Finance*, 1987, *42*, 965–986.

Bradley, Michael, Amand Desai, and E. Han Kim, "Determinants of the Wealth Effects of Corporate Acquisitions," working paper, The University of Michigan, 1984.

Bradley, Michael, Amand Desai, and E. Han Kim, "The Rationale Behind Interfirm Tender Offers: Information or Synergy?" *Journal of Financial Economics*, 1983, *11*, 183–206.

Brickley, James, Ronald Lease, and Clifford Smith, "Ownership Structure and the Voting on Antitakeover Amendments," *Journal of Financial Economics*, forthcoming.

Brown, Charles and James L. Medoff, "The Impact of Firm Acquisitions on Labor," NBER Working paper, 1987.

Brown, Stephen J. and Jerold B. Warner, "Using Daily Stock Returns: The Case of Event Studies," *Journal of Financial Economics*, 1985, *14*, 3–31.

Comment, Robert, "Price and Volume Before Tender Offers: Market Anticipation Activity or Inside Trading, working paper New York University, 1986.

Comment, Robert and Gregg A. Jarrell, "Two-Tier Tender Offers: The Imprisonment of the Free Riding Shareholder," *Journal of Financial Economics*, forthcoming.

Council of Economic Advisors, "The Market for Corporate Control," *Economic Report of the President*, 1985, 187–216.

DeAngelo, Harry, and Linda DeAngelo, "Managerial Ownership of Voting Rights: A Study of Public Corporations With Dual Classes of Common Stock," *Journal of Financial Economics*, 1985, *14*, 33–69.

DeAngelo, Harry, Linda DeAngelo, and Edward M. Rice, "Going Private: Minority Freezeouts and Stockholder Wealth," *Journal of Law and Economics*, 1984, *27*, 367–402.

Denis, Debra K., and John J. McConnell, "Corporate Mergers and Security Returns," *Journal of Financial Economics*, 1986, *16*, 143–187.

Dodd, Peter, and Richard Leftwich, "The Market for Corporate Charters: 'Unhealthy Competition' Versus Federal Regulation," *Journal of Business*, 1980, *53*, 259–283.

Easterbrook, Frank H., and Gregg A. Jarrell, "Do Targets Gain From Defeating Tender Offers?" *New York University Law Review*, 1984, *59*, 277–299.

Eckbo, B. Espen, "Merger and the Market Concentration Doctrine: Evidence from the Capital Market," *Journal of Business*, 1985, *58*, 325–349.

Eckbo, B. Espen, and Herwig Langohr, "The Effect of Disclosure Regulations and the Medium of Exchange on Takeover Bids," Working paper, 1986.

Guerin-Calvert, Margaret, Robert H. McGuckin, and Frederick R. Warren-Boulton, "State and Federal Regulation in the Market for Corporate Control," U.S. Department of Justice, Economic Analysis Group Discussion Paper 86-4, 1986.

Hall, Bronwyn H., "The Effect of Takeover Activity on Corporate Research and Development," working paper, NBER, 1987.

Holderness, Clifford G. and Dennis P. Sheehan, "Raiders or Saviors? The Evidence on Six Controversial Investors," *Journal of Financial Economics*, 1985, *14*, 555–579.

James, Christopher M., and Peggy Weir, "Determinants of the Division of Gains in Corporate Acquisitions: Evidence from the Banking Industry, working paper, University of Oregon, 1987.

Jarrell, Gregg A., "The Wealth Effects of Litigation by Targets: Do Interests Diverge in a Merge?" *Journal of Law and Economics*, 1985, *28*, 151–177.

Jarrell, Gregg A., and Michael Bradley, "The Economic Effects of Federal and State Regulations of Cash Tender Offers," *Journal of Law and Economics*, 1980, *23*, 371–407.

Jarrell, Gregg A., and Annette B. Poulsen, "Bidder Returns," working paper, 1987a.

Jarrell, Gregg A., and Annette B. Poulsen, "The Effects of Recapitalization with Dual Classes of Common Stock on the Wealth of Shareholders," *Journal of Financial Economics*, forthcoming.

Jarrell, Gregg A., and Annette B. Poulsen, "Shark Repellents and Stock Prices: The Effects of Antitakeover Amendments Since 1980," *Journal of Financial Economics*, 1987b, *19*, 127–168.

Jensen, Michael C., "The Takeover Controversy: Analysis and Evidence," *Midland Corporate Finance Journal*, 1986, 6–32.

Jensen, Michael C., and Richard S. Ruback, "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics*, 1983, *11*, 5–50.

Lease, Ronald C., John J. McConnell, and Wayne H. Mikkelson, "The Market Value of Control in Publicly Traded Corporations," *Journal of Financial Economics*, 1983, *11*, 439–472.

Lehn, Kenneth, and Annette B. Poulsen, "Sources of Value in Leveraged Buyouts." In *Public Policy Towards Corporate Takeovers*. New Brunswick, NJ: Transaction Publishers, 1987.

Malatesta, Paul H., and Ralph A. Walkling, "Poison Pill Securities: Stockholder Wealth, Profitability, and Ownership Structure," *Journal of Financial Economics*, forthcoming.

McConnell, John J. and Chris J. Muscarella, "Capital Expenditure Decisions and Market Value of the Firm," *Journal of Financial Economics*, 1985, *14*, 399–422.

Mikkelson, Wayne H. and Richard S. Ruback, "An Empirical Analysis of the Interfirm Equity Investment Process," *Journal of Financial Economics*, 1985, *14*, 523–553.

Netter, Jeffry M. "Shareholder Wealth Effects of Litigation Based on Allegedly False Schedule 13D Item 4 Disclosure," working paper, 1987.

Office of the Chief Economist, Securities and Exchange Commission, "The Economics of Any-or-All, Partial, and Two-Tier Tender Offers," 1985a.

Office of the Chief Economist, Securities and Exchange Commission, "Update—The Effects of Dual-Class Recapitalizations on Shareholder Wealth: Including Evidence from 1986 and 1987," 1987d.

Office of the Chief Economist, Securities and Exchange Commission, "The Effects of Poison Pills on the Wealth of Target Shareholders," 1986.

Office of the Chief Economist, Securities and Exchange Commission, "The Impact of Targetted Share Repurchases (Greenmail) on Stock Prices," 1984.

Office of the Chief Economist, Securities and Exchange Commission, "Institutional Ownership, Tender Offers and Long Term Investment," 1985b.

Office of the Chief Economist, Securities and Exchange Commission, "Shareholder Wealth Effects of Ohio Legislation Affecting Takeovers," 1987b.

Office of the Chief Economist, Securities and Exchange Commission, "Shark Repellents and Stock Prices: The Effects of Antitakeover Amendments Since 1980," 1985c.

Office of the Chief Economist, Securities and Exchange Commission, "Stock Trading Before the Announcement of Tender Offers: Insider Trading or Market Anticipation?" 1987c.

Office of the Chief Economist, Securities and Exchange Commission, "The Effects of Dual-Class Recapitalizations on the Wealth Of Shareholders," 1987a.

Partch, Megan. "The Creation of A Class of Limited Voting Common Stock and Shareholders" Wealth," *Journal of Financial Economics*, 1987, *18*, 313–339.

Romano, Roberta, "Law as a Product: Some Pieces of the Incorporation Puzzle," *Journal of Law Economics and Organization*, 1985, *1*, 225–267.

Ruback, Richard S., "An Overview of Takeover Defenses," working paper #1836-86, Massachusetts Institute of Technology, 1986.

Ryngaert, Michael, "The Effect of Poison Pill Securities on Shareholder Wealth," *Journal of Financial Economics*, forthcoming.

Ryngaert, Michael, and Nettter Jeffry, "Shareholder Wealth Effects of the Ohio Antitakeover Law," working paper, 1987.

Schumann, Laurence, "State Regulation of Takeovers and Shareholder Wealth: The Effects of New York's 1985 Takeover Statutes," Bureau of Economics Staff Report to the Federal Trade Commission, March 1987.

Shleifer, Andrei and Lawrence Summers, "Hostile Takeovers as Breaches of Trust," working paper, NBER, 1987.

Smiley, Robert, "The Effect of State Securities Statutes on Tender Offer Activity," *Economic Inquiry*, 1981, *19*, 426–435.

Steindel, Charles, "Tax Reform and the Merger and Acquisition Market: The Repeal of General Utilities," *Federal Reserve Bulletin of New York*, Autumn 1986, 31–35.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 69–82

Corporate Takeovers: The Efficiency Arguments

F. M. Scherer

In recent years, the tender offer takeover has been praised and damned with a ferocity suggesting that the survival of capitalism is at stake. The truth, as in most disputes with substantial metaphysical content, is more prosaic. Some takeovers enhance economic efficiency, some degrade it, and the balance of effects, though not fully known, is most likely a close one. In this essay I try to lay bare the debate's foundations and bring it back to earth with an injection of evidence.

The Theoretical Bases

No match made in heaven is more blissful than an extant economic theory that finds an important real-world phenomenon to explain. The theory of takeovers articulated in Robin Marris's seminal article (1963) argues that two market failures can be corrected by the proper functioning of a third market—the market for corporate control. Competition in product and input markets may fail, it is said, to weed out firms that have strayed from the path of cost minimization and profit maximization. Within companies free from strong external market pressure, the "separation of ownership and control" identified by Berle and Means (1932) allows managers to pursue investment and cost-padding strategies that fail to maximize the profits of absentee owners—the shareholders. But a well-functioning market for corporate control comes to the rescue. Outside interests, seeing that profits would be

■ *F. M. Scherer is Joseph Wharton Professor of Political Economy, Swarthmore College, Swarthmore, Pennsylvania.*

higher if different strategic choices were made, bid on the stock market for a controlling stake in the firm, take the company over, enforce new profit-maximizing policies, and live happily (efficiently) ever after. The theory was prescient, because at the time it was articulated, the tender offer takeover was a rare anomaly on the capitalist scene.[1] But in the late 1960s and much more so during the 1980s, tender offer takeover activity has grown to dramatic levels. The result has been a renaissance of interest in the theory of takeovers.

Nevertheless, viewing the theory's applicability from a broad historical perspective raises awkward questions. The "separation of ownership and control," though increasing over time, was found by Berle and Means to be widespread among large U.S. corporations during the 1920s. If managerial deviations from profit maximization were already commonplace then, how can one explain the strong performance of the U.S. industrial economy during the 1950s and 1960s, before efficiency-restoring tender offer takeovers came into vogue? Also, the tender offer takeover phenomenon is both new and almost uniquely Anglo-American. If takeovers are necessary for efficiency, how have nations such as Japan, West Germany, Switzerland, and France continued to perform strongly, even though hostile takeovers are practically nonexistent there, and stock ownership is often separated from managerial control, as in America and Great Britain? The questions are rhetorical, for no satisfactory answer has been forthcoming.

Inherent Limitations of Event Study Evidence

Much of the empirical support for takeovers as an efficiency-increasing force comes from a new genre of statistical research. In 1974, Gershon Mandelker published the first merger "event study" rooted in the logic of the Capital Asset Pricing Model. Since then, the Center for Research on Security Prices has made available machine-readable data on daily stock price movements for a large population of corporations. The consequence of these two innovations has been the economics and corporate finance profession's closest analogue to the automated factory. Scores of papers have been written on how the announcement of a merger, takeover or related corporate control transaction affects the normalized stock prices of target and acquiring firms.

As the paper by Jarrell, Brickley and Netter in this symposium reveals, these merger event studies have yielded impressive results. Within a several-week time "window" around the announcement of a merger, whether voluntary or through takeover, stockholders of the target company experience substantial positive abnormal returns—that is, gains beyond those accruing had the target's stock price merely

[1] Although rare, takeovers did occur. Marris (1963, p. 189) notes that his views were influenced by "the dramatic crop of raids, both successful and unsuccessful, which has occurred in the United Kingdom in the last ten years."

tracked overall market movements. For acquiring firms the results are less uniform, but the most common finding has been that no significant abnormal returns, positive or negative, appear. From the positive returns to targets and the small or zero returns to acquirers, merger event analysts generally conclude that mergers are on average value-enhancing. With some further assumptions, only two of which will be examined critically here, the conclusion is drawn that mergers are also efficiency-enhancing. The presumed efficiencies may stem from the displacement of inefficient managers, a reorientation of company business strategy, or the realization of economies of scale or scope. However, for a variety of reasons, there are serious gaps in the chain of logic leading from observed stock price increases to the inference that economic efficiency is improved.

Long-Term Losses for Acquiring Firms

Most merger event studies focus on stock price movements during only a few weeks surrounding the relevant announcement. When the time frame has been extended to one to three years after the event, acquiring firms are found to experience *negative* abnormal returns. In the seven one-year studies surveyed by Jensen and Ruback (1983), the abnormal returns averaged −5.5 percent; over the three-year post-takeover period examined by Magenheim and Mueller (1987), the abnormal returns were −16 percent by the most conservative measurement technique.[2] These negative post-merger abnormal returns often turn out to be statistically insignificant, since as is common in unbounded random walk problems, the variance rises with the length of the time period studied. Yet as Jensen and Ruback concede in their review of the evidence (p. 20), "These post-outcome negative abnormal returns are unsettling because they are inconsistent with market efficiency and suggest that changes in stock price during takeovers overestimate the future efficiency gains from merger."

A still longer-term perspective provides further grounds for skepticism. Early event study samples emphasized the preponderantly conglomerate mergers of the 1960s and 1970s. As in the newer studies, positive short-term stock market reactions were observed. Yet with two decades of hindsight, there is now considerable agreement that the conglomerate merger-making peaking in 1968 led to widespread failure, evidenced in low returns to conglomerate firms' shareholders and extensive divestiture of ill-fitting, poorly-managed subsidiaries.[3] The implications of event studies for those early mergers have been contradicted by historical evidence. A technique that fails once may work on subsequent iterations. But the failure in an important test at least advises caution in interpreting event studies on mergers of the 1980s.

---

[2] It is worth noting that when Ivan Boesky obtained advance information on an impending takeover, he often sold the shares of the acquirer short. See "Boesky Apparently Reaped at Least $203 Million in Illicit Profits with Levine's Inside Information," *Wall Street Journal*, November 24, 1986, p. 2.

[3] Scholarly disagreement on this point, once strong, appears to be waning. Compare Ravenscraft and Scherer (1987c), Chapters 7 and 8, with the statement of Michael Jensen in U.S. House of Representatives (1987), p. 137, in which managers with large free cash flow and/or borrowing power are said to engage in "low-benefit or even value-destroying" diversification mergers. See also footnote 10.

Market Efficiency as an Operational Concept

Merger event studies usually accept the stock market's efficiency—that is, the tendency for stock prices to impound all available information at any moment in time—as an axiom, not as a hypothesis to be tested. Indeed, empirical proof or refutation is singularly difficult (Summers, 1986). Differential willingness to accept the efficient markets axiom is at the heart of much disputation between those who infer from event studies that takeovers are efficiency-enhancing and those who are more skeptical.

The skeptical view of takeovers advances an alternative explanation for the stock price behavior observed in merger event studies. It is well accepted that stock prices move over time in something approximating a random walk. Companies can therefore become targets not only because their managers have erred in failing to maximize profits, but because the stock market has erred, randomly (or through fad-like movements (Shiller, 1987)), setting share prices so low as to make their issuer a bargain worth snapping up.[4] Random stock market valuation errors may also nominate likely acquirers, for the company whose stock is overvalued (and knows it) has a uniquely economical currency with which to make acquisitions or raise the cash for making them. Thus, premiums may be paid to gain control of undervalued companies even when no efficiency gains are expected to result from the ownership change.

On both practical and theoretical grounds, it appears likely that realistically efficient stock markets generate many opportunities for takeovers motivated by the perception that companies are undervalued. In his presidential address to the American Finance Association, Fischer Black (1986, p. 533) framed market efficiency in concrete operational terms:

> [W]e might define an efficient market as one in which price is within a factor of 2 of value, i.e., the price is more than half of value and less than twice value. The factor of 2 is arbitrary, of course. Intuitively, though, it seems reasonable to me, in the light of sources of uncertainty about value and the strength of the forces tending to cause price to return to value. By this definition, I think almost all markets are efficient almost all of the time. "Almost all" means at least 90 percent.

If Black's estimate represents the 90 percent confidence bounds about a log normal distribution, for example, then 16 percent of corporate stocks would be undervalued by 34 percent or more at any time. Such a distribution of actual prices creates enormous incentives for would-be acquirers who believe that their estimate of true value is more accurate, or based upon superior insider information, than the stock market's.

---

[4] Merger-makers often behave, or at least claim to behave, as if the efficiency axiom were untrue. They search actively for "bargain" companies to acquire. For numerous examples, see Ravenscraft and Scherer (1987c), pp. 9–10.

The sources of the market's valuation errors are not made clear by Black. One important element, as Shiller (1981) and others have argued, must come from exaggerated general market fluctuations over time. Consider Mobil Corporation's incentive in 1981 to take over Marathon Oil.[5] The day before Mobil's first $85 per share tender offer, Marathon's stock closed at $63.75. What especially attracted Mobil to attempt a takeover was Marathon's 49 percent interest in the Yates oil field, with estimated reserves sufficient to support low-cost production of 22.4 million barrels per year for 45 years. Assuming (as most analysts did at the time) continuation of the prevailing $35 per barrel price, and applying the average 16.04 percent Baa bond rate of 1981, the discounted present value of Yates field oil is found to be $4.74 billion —slightly less than Mobil's first bid for all of Marathon's assets.[6] But suppose one believed that the tight monetary conditions of 1981 could not persist, so one correctly forecasted that in 1986, Baa bonds could be floated at a 10.39 percent average rate. Assuming that bridge financing was secured at the 16.04 percent rate of 1981, and the 1986 rate applied until the field was depleted in the year 2026, the discounted present value would have been $6.90 billion—46 percent more than the value assuming 1981 rates to persist indefinitely. Thus, to an acquirer with not unreasonable foresight about interest rate movements (even if not about oil prices), companies whose future cash flows were discounted heavily by the market in 1981 were attractive bargains. I believe many takeovers of the early 1980s had an explanation as simple as this. Speculators who paid a 30 to 40 percent or so stock price premium to acquire companies when the Standard & Poor's 500 company index stood at 128 in 1981 or 120 in 1982 could feel well satisfied with their purchases as the index moved to an average 1986 level of 236—especially when the gamble was made largely with other shareholders' or banks' money.

The most persuasive argument against such value error motives for merger comes from studies of stock price behavior following takeover attempts that failed. (Again, Jarrell, Brickley and Netter describe the evidence in this symposium.) After a tender offer fails, target company share values characteristically remain for a while near the higher level to which they were bid at the initial takeover announcement. Once a company is "in play," there is a sizable probability that a subsequent offer will materialize, and if it succeeds, the persisting positive abnormal return is validated. But

[5] The author was a witness for Marathon Oil in the ensuing antitrust case.

[6] Shortly after the tender offer was made, Mobil's president testified that his finance staff had estimated the value of Marathon's U.S. crude oil reserves, including several fields in addition to Yates, to be "as high as $180" per share, or $10.9 billion in total. *Marathon Oil Co. v. Mobil Corporation et al.*, N. Dist. Ohio (1981), hearing transcript p. 448. The present values presented here are computed by applying to annual cash inflows the factor $[e^{-rI} - e^{-rT}]/r$, where r is the applicable discount rate, I is the number of years hence at which the calculation begins, and T is the number of years hence at which the calculation ends.

if no new takeover occurs within two years, target stock prices tend to retreat toward their pre-first-offer level. Event analysts interpret this evidence as showing that value is enhanced only when a takeover actually occurs, with its accompanying internal reorganizations, and not by the signalling effect of an unsuccessful takeover.

An equally plausible but more skeptical interpretation is that some tender offer targets are not truly undervalued, while others are. During the "in play" period, targets are subjected to intense scrutiny. Those that were truly undervalued but escape the first tenderer will prove attractive to another acquirer, while the small minority that were not undervalued will fall through the screen and see their share values return to the (correct) pre-announcement levels. The evidence from event studies is incapable of distinguishing between these contending hypotheses.

Line of Business Data: An Alternative Approach

If merger event studies invoking the efficient markets axiom cannot demonstrate whether mergers are in fact efficiency-enhancing, what other recourse is there? One alternative is to investigate the internal financial health of companies at the time they become takeover targets and to ascertain how their profit performance changes over a substantial period of time following merger. The corporate control debate suggests two hypotheses testable under this approach. First, if tender offers are precipitated by incumbent management failures, tender offer targets should be less profitable than peer companies in similar industry groups. Second, if takeover displaces inefficient management or in other ways facilitates a movement to higher profit business strategies, post-takeover profitability should rise relative to pre-takeover profitability, conditions in the acquired unit's home industry being held constant. David Ravenscraft and I (1987b) have carried out such an analysis using unprecedentedly detailed Line of Business data collected by the Federal Trade Commission.

For the task at hand, Line of Business data has two compelling advantages. With parent company operations disaggregated into detailed industry categories (for manufacturing, encompassing 261 distinct possibilities), acquired units can be singled out for special scrutiny. Their sales, assets, and profits are much less likely to be swamped by the larger magnitudes from other company activities. Also, with data disaggregated to the narrow industry category level, it is possible to form well-focused controls—notably, the performance of other companies' non-acquired units of comparable size in the same industry.

The FTC's Line of Business surveys covered a maximum of 471 U.S. industrial corporations that made three-fourths of all manufacturing and mineral company acquisitions, by asset value, between 1950 and 1976. To the 4,409 individual manufacturing lines of business on which the sample companies filed one or more annual reports, nearly 6,000 acquisitions consummated between 1950 and 1976 were coded. Of these, 96 were either tender offer takeovers or "white knight" transactions triggered by a tender offer. Given the special interest in the efficiency implications of tender offers, I focus here on that subset of our acquired unit sample.[7]

For the 95 tender offer targets on which pre-takeover data were available, operating income (before deduction of interest charges, income taxes, and extraordinary items) averaged 11.08 percent of assets. The time-matched average for the principal two-digit manufacturing industries in which the tender offer targets operated was 12.06 percent. The difference is significant at the 0.05 level (t = 1.90). Thus, on average, the tender offer targets were under-performers by about 8 percent relative to their home industry norms.[8]

Post-takeover performance was analyzed by averaging profitability indices of the acquired lines for the three years 1975–77, which on average followed the takeover by nine years. Lines subjected to tender offer-induced acquisition were 23 percent less profitable on average than otherwise comparable lines not involved in tender offers, holding constant (among other things) the 21 percent degradation of profitability associated with the use of purchase accounting to value acquired assets.[9]

The accounting method used in valuing acquired assets is important in interpreting these results. Under purchase accounting, accepted for most tender offer takeovers as well as many other acquisitions, assets are written up to reflect the actual amounts paid to acquire them. If a premium over premerger book value is paid, as is typical in tender offer acquisitions, the denominator of the ratio between operating income and assets is inflated relative to the denominators of nonmerged units retaining their book values unchanged. With larger asset values, numerators of the ratio may also be reduced owing to higher postmerger depreciation charges. To filter out these accounting effects stemming from the payment of takeover premiums, the ratio of cash flow (before deduction of depreciation) to sales was substituted as dependent variable in an otherwise identical analysis. The estimated tender offer effect was still negative but smaller and statistically insignificant, showing average tender offer line performance 11 percent less than that of otherwise comparable non-tender lines. This implies that much of the observed deterioration of the ratio of operating income to assets associated with a tender offer history came from paying a high price to make the takeover and restating asset values to reflect that price.

[7] Results for the much larger voluntary merger group are reported in Ravenscraft and Scherer (1987c).

[8] For 56 targets of tender offers initiated between 1975 and 1983—a sample more recent than ours—Herman and Lowenstein (1987) found that returns on equity were superior to those of all manufacturing corporations when a weighted average was taken. By my reworking of their reported results, the simple average returns appear to have been slightly below the 1975–83 all-manufacturing average. Evidently, the largest targets—frequently oil companies—had profits well above average.

[9] A fixed effects multiple regression model was used. For 2,732 lines of business with high-quality data for the three years, profitability indices were regressed on a tender offer dummy variable (with unit value for the 153 lines with a tender offer history), a variable measuring the fraction of assets acquired by the parent company under purchase accounting, a variable measuring the line's market share, 256 industry intercept dummy variables, and three other control variables of less interest to the present discussion. In the regression of operating income as a percent of assets on these variables, the tender offer dummy variable had a value of −3.10, with t = 2.41. The purchase accounting coefficient was −2.81, with t = 2.24. Full-sample mean profitability was 13.3 percent.

Since the acquired companies had slightly inferior profit performance before takeover and, abstracting from accounting revaluation effects, the acquired lines continued to have slightly inferior cash flow/sales performance after takeover, one must conclude that operating performance neither improved nor deteriorated significantly following takeover. The hypothesis that takeovers improve performance is not supported.

Special Characteristics of the 1980s Takeover Wave

These results are for a data set covering takeovers up to the mid-1970s. It is uncertain whether the statistical results will extrapolate to tender offer takeovers of the 1980s, which are different from their predecessors in numerous respects. The new takeovers are larger on average and more likely to be made by blue chip corporations rather than maverick "raider" types. They have several other noteworthy characteristics.

"Bust Up" Takeovers

The (mostly voluntary) merger wave of the 1960s and early 1970s was preponderantly conglomerate, more than doubling the number of lines in which the average Line of Business survey company operated. By contrast, the 1980s have seen a high incidence of "bust up" takeovers—that is, acquisitions followed by the sell-off of numerous target company divisions. Divestiture in such cases has been motivated by two main considerations—a need to retire loans incurred in making the takeover, and a belief that the parts individually were worth more than the whole. To reduce debt, on which I shall say more shortly, "crown jewel" divisions are sometimes sold. The "parts worth more than the whole" motive is more interesting. The research by Ravenscraft and myself (1987c, Chapters 5 and 6) sheds new light on it.

With few exceptions, the diversifying acquisitions of the 1960s and 1970s were much less than a resounding success. For acquired lines surviving long enough to be included in the 1975–77 Line of Business surveys, profitability fell sharply on average relative to premerger levels. Moreover, Ravenscraft and I estimate conservatively that by 1981, one-third of the units acquired had been sold off.[10] On average, lines that were fully divested had *negative* operating income in the year before sell-off commenced—a clear sign of failure. Fifteen case studies of acquired-and-then-divested units revealed that sell-off was often precipitated by managerial control loss and incentive breakdowns. These in turn had roots in the more complex organizational structures into which the acquired units were thrust, knowledge lacunae that impaired the conglomerate parent's ability to solve emerging problems, and the inability of top management to develop incentives stimulating sustained, vigorous performance by unit operating heads. When sell-off occurred, the divested units typically moved from a conglomerate organizational structure to either a much simpler structure (under unit leveraged buyouts),[11] or to a horizontal or vertical relationship in which their new parent had substantial experience relevant to the divested-and-reacquired unit's business. Performance improvements, evident both qualitatively (from the case studies) and quantitatively (for units sold off to Line of Business sample parents) followed.

---

[10] Porter (1987, p. 48) finds an even higher 56.5 percent sell-off rate for pre-1976 acquisitions of 33 diversification-prone U.S. corporations and concludes, "The corporate strategies of most companies have dissipated instead of created shareholder value."

The cumulation of these developments has become known as the "back to basics" movement in American industry. "Bust up" takeovers have played a role. Managements that for some reason were unwilling to divest units that could be more profitable in another organizational structure have been forced to do so, either by displacement or to preempt feared takeovers. In this way, takeovers have been efficiency-increasing. But one should not exaggerate their role. The number of divisional sell-offs by U.S. corporations reached a peak in 1971, three years after the peak of the conglomerate merger wave and long before the "bust up" takeover became known. After declining by 65 percent relative to the peak, the number of sell-offs began rising again from 1981 on. "Bust up" takeovers undoubtedly contributed to the increase, but they were not yet popular in 1981, when the turnaround began. Another contributor was the increase in friendly merger activity, creating more ill fits that needed undoing.

A statistical analysis of Line of Business sample unit sell-offs between 1975 and 1981 showed that the probability of divestiture moved from zero to near unity with four standard deviation changes in each of six variables: line of business profitability, company-wide profitability, the degree to which a subsidiary was related conglomerately to its parent, the line's market share, research and development intensity, and change in the identity of the company's chief executive officer (Ravenscraft and Scherer, 1987a).[12] Takeovers may have raised the probabilities in the 1980s, but they cannot have raised them much. If a unit's performance was seriously deficient, sell-off was likely to occur in any event, especially after the parent's chief executive officer (often responsible for the earlier acquisition mistake) was replaced. (On average, our research revealed, such CEO changes occurred every eleven years.) Thus, much efficiency-enhancing sell-off activity was precipitated by such fundamental variables as profitability, strategic position, and "fit." Takeovers and their threat contributed only at the margin.

Debt Financing

Another prominent characteristic of takeovers in the 1980s is the extensive use of debt financing, introducing increased leverage into the merged company's financial structure.[13] Potential target companies have also been induced to refinance and accept increased leverage to deter takeover attempts. From the perspective of individual firms, debt is advantageous relative to common stock because its interest is deductible in computing taxable corporate income, avoiding the double taxation imposed upon dividends.[14] Within limits, more debt financing reduces a corporation's cost of capital. Lower capital costs may in turn induce more growth-creating, productivity-raising real capital investment and tilt investment decisions in more far-sighted directions. In this respect, the greater reliance upon debt stimulated in part by the takeover movement is efficiency-enhancing, correcting the cost of capital disparities between U.S. corporations and their overseas rivals that became increasingly problematic as international competition intensified during the 1980s. The limits beyond which the risks of leverage, and hence risk premiums, rise sharply do not appear thus far to have been seriously overstepped. For the most part, the financial markets have reacted favorably.

---

[11] Under a unit leveraged buyout, the sold-off unit becomes a new free standing enterprise, usually with a heavily debt leveraged capital structure. Heavy borrowing permits the unit's management to hold a substantial share of the outstanding common stock shares.

[12] In a logit analysis with seven other variables held constant at their means, each of the six was varied symmetrically about its own mean by four standard deviations in a sell-off probability-enhancing direction, and the predicted change in the sell-off probability for these combined changes was evaluated.

But heavy leveraging imposes costs external to the individual debt-issuing firm and its creditors. As the replacement of equity with debt spreads to a substantial fraction of all industry, the macroeconomy becomes more unstable. If a recession occurs, highly leveraged companies whose cash flow has fallen may be forced to cut back real investment in order to service their debt and avoid default.[15] The larger real investment cutbacks will be augmented by conventional autonomous investment multipliers, accentuating the magnitude of the downturn. This external cost of increased leverage could be serious, especially when the ability of the federal government to implement countercyclical measures is constrained by concern over the size of its budget deficits. Thus, the financial structure changes associated with takeover activity in the 1980s are a mixed blessing, meriting at best two cheers, and weak ones at that, on broad economic efficiency grounds.

Short-Term Managerial Horizons

Perhaps the most controversial and least understood consequence of the 1980s takeover movement is its alleged propensity to shorten managerial decision-making horizons and bias profit maximization toward the short run. If the allegation is true, it could be the most important implication of takeovers, since American industry is confronted by European and especially Asian rivals who invest heavily and practice penetration and learning curve pricing[16]—manifestations of a long-run strategy. To see how takeovers affect managerial time horizons and are affected by them, it is useful to distinguish between the early and late phases of the recent merger wave.

---

[13] "Funny money"—often, convertible preferred stock—was also widely used to finance conglomerate acquisitions of the 1960s.

[14] Another advantage emphasized by Jensen in this symposium is that debt forces companies to pay out cash flows that would otherwise be invested unwisely. The excess cash flow idea was proposed and tested earlier by Mueller (1969) and Grabowski and Mueller (1975).

[15] Galbraith (1955, p. 183) notes a similar phenomenon as one contributor to the severity of the Great Crash in 1929.

[16] On learning curve pricing, which entails setting low prices at early stages in a product's life to advance more rapidly down the learning curve and gain cost advantages over rivals, see Spence (1981) and Ross (1986).

From 1979 to 1982, both nominal and real interest rates rose sharply, signalling that capital investment should be reduced and reoriented toward projects with shorter-run cash return profiles. But there is reason to believe that large corporations adjust the "hurdle rates" applied in capital budgeting decisions only with a lag, in effect "smoothing" out fluctuations in underlying capital market conditions (Hayes, 1977; Donaldson and Lorsch, 1983, pp. 51–53, 169–171). Data from the two most closely comparable surveys indicate that "hurdle rates" (computed mostly on an after-tax basis) rose from a median value of 12 percent in 1969 to 15 percent in June 1983.[17] Meanwhile, after-tax corporate capital costs probably rose by at least 5 percentage points, excluding risk premium adjustments. Companies whose hurdle rates lagged during the early 1980s interest rate upsurge were apt to approve more far-sighted investment projects than what the market was demanding. This disparity could have made the companies vulnerable to acquirers who, following takeover, would enforce investment policies more to the market's liking. Fear of being displaced in this way probably underlay early top management complaints that takeover threats were forcing their firms to become excessively short-sighted. If interest rates were more volatile than long-run conditions warranted, it is by no means clear that managers were being inefficient in their rate-smoothing behavior, therefore meriting displacement through takeovers.

Since 1983, nominal interest rates have fallen, so dissonance between rate-smoothing firms and the market seems likely to have abated. If a problem remains, it must have other more subtle roots. Business leaders claim that "the market," and especially high-turnover stock fund managers, are excessively short-sighted, but clear-cut behavioral consequences have not been identified. The most pointed suggestions come from National Science Foundation, Battelle Institute, and McGraw Hill surveys reporting sharply lower company-financed research and development spending growth rates in 1986 and 1987, and attributing the changes partly to profit-raising efforts induced by takeover fears.[18] All three sources concede that the causal link between takeover threats and R&D decisions is ambiguous. On the more optimistic side, a Securities and Exchange Commission study (1985) found that institutional investors and the stock market responded positively to R&D increases and research-intensive corporations generally. However, the samples used in that study were small or biased toward high R&D companies and perhaps toward the venture capital funds investing in such companies. The analyses suffered from a lack of industry controls, which earlier research has shown to be crucial if valid conclusions are to be drawn. More equivocal results on the R&D impact of 1976–85 mergers, voluntary and hostile, are reported in Hall (1987). In the present state of knowledge, the question of how takeovers affect R&D spending and other far-sighted investment remains unresolved. Analyses using richer data with superior controls and more precise identification of lagged behavioral effects are needed. Alternatively, new insights might be gained through carefully structured interviews with a large cohort of business decision-makers. Until such progress occurs, we remain in the dark on a potentially important efficiency effect of takeovers.

---

[17] Compare Williams (1970, p. 21) and Scherer (1983). My own survey, with usable responses from 42 *Fortune* 500 corporations, revealed an anomaly. Half the companies adjusted future cash flows to reflect expected inflation, while the other half did not. The former should have used nominal interest rates in computing their hurdle rates, while the latter should have used real interest rates. The difference between their hurdle rates was in the expected direction, but averaged only 2.72 percentage points. This surprisingly small difference may imply either misunderstanding of rather clear survey questions, or failure to understand and/or apply the relevant economic theory. If the latter, the non-inflation-adjusters probably biased their decisions too much toward the short run. But even so, stickiness in the face of 1979–82 interest rate increases implies a marginal lengthening of time horizons relative to what the capital markets signalled.

Less comparable surveys by Hayes (1977), Gitman and Forrester (1977), and Gitman and Mercurio (1982) yield *mean* (not median) hurdle, cost of capital or "cut off" rates of 13.9 percent in 1975, 14.0 percent in June 1976, and 14.3 percent in October 1980. Again, the smoothing implication is supported.

[18] "The Sun May Be Setting on R&D's Glory Days," *Business Week*, January 19, 1987, p. 75; "Growth in Company-Funded Research and Development Expected to Slow in 1987," National Science Foundation, *Science Resources Studies Highlights*, NSF 86-314, January 30, 1987; and "Corporations Are Putting Less into Research," *Business Week*, August 10, 1987, p. 18.

Conclusion

In theory, tender offer takeovers provide a significant corrective against managerial departures from profit maximization. Careful scrutiny of the available evidence leads to a more skeptical assessment. The market processes that put a company "in play" are at best noisy, creating incentives for takeover both when management errs and when the market errs. Consistent with this mix of incentives, target companies are seen on average to be only mild profit under-performers. Evidence on 1960s and early 1970s tender offer takeovers reveals no significant long-term improvement in operating profitability following takeover. Recent takeovers have probably accelerated needed divisional restructuring and led to lower-cost financial structures, but not without risks to the macroeconomy. Whether such takeovers and the more pervasive threat that they will occur have wider-ranging behavioral consequences remains unknown, leaving important challenges for future research.

As this paper was being written, numerous proposals for change in the laws and regulations governing takeover activity were pending before the U.S. Congress and state legislatures. Some of the proposals, by intent or as a by-product of "fairness" or procedural transparency goals, would make it more difficult to effect tender offer takeovers. The most vigorous principled objections to such changes are rooted in a belief that takeovers and their threat perform an important efficiency-enhancing function in the U.S. economy. In this paper, I have questioned both the theoretical and empirical bases for that belief. If takeovers on average do little to improve corporations' operating efficiency, the objection to regulations emphasizing fairness and other criteria loses much of its force, and the case for slowing down what is clearly a costly and disruptive corporate restructuring process gains plausibility.

On the other hand, should the supporters of wide-open corporate control markets be correct in their belief that the separation of ownership and control causes serious inefficiencies, one must ask whether remedies simpler and more direct than the takeover mechanism might exist. If the real problem is that managers ride roughshod over absentee shareholders' preferences, a more direct solution would be to ensure that shareholders are better represented on more active corporate boards. Louis Lowenstein (forthcoming) has proposed, for example, that a sizable fraction of public corporations' directors be nominated directly by nonmanagement owners. Since roughly half of the typical large corporation's shares are held by financial intermediaries—including pension funds (Scherer, forthcoming)—this means in effect that significant nominating powers would be vested in the managers and trustees of such large funds as the College Retirement Equities Fund. Economist-trustees of CREF were instrumental in redirecting its investment policies away from costly "churning" strategies and toward practices consistent with the teachings of modern finance theory. It is an interesting question whether they might now help reform the target corporations themselves, if reform is in fact needed, by nominating superior "outside" directors. I support such an experiment in improved corporate governance much more readily than I can support the radical wide-open tender offer experiments, which place corporate management at risk whether it, or the stock market, has been guilty of error.

References

Berle, Adolf A., Jr., and Gardiner C. Means, *The Modern Corporation and Private Property.* Chicago: Commerce Clearing House, 1932.

Black, Fischer, "Noise," *Journal of Finance,* July 1986, *41,* 529–543.

Donaldson, Gordon, and Jay W. Lorsch, *Decision Making at the Top.* New York: Basic Books, 1983.

Galbraith, John Kenneth, *The Great Crash.* Boston: Houghton-Mifflin, 1955.

Gitman, Lawrence J., and John R. Forrester, Jr., "A Survey of Capital Budgeting Techniques Used by Major U.S. Firms," *Financial Management,* Fall 1977, *6,* 66–71.

Gitman, Lawrence J., and Vincent A. Mercurio, "Cost of Capital Techniques Used by Major U.S. Firms," *Financial Management,* Winter 1982, *11,* 21–29.

Grabowski, Henry, and Dennis C. Mueller, "Life-Cycle Effects on Corporate Returns on Retentions," *Review of Economics and Statistics,* November 1975, *57,* 400–409.

Hall, Bronwyn H., "The Effect of Takeover Activity on Corporate Research and Development," National Bureau of Economic Research Working Paper No. 2191, March 1987.

Hayes, Samuel L., "Capital Commitments and the High Cost of Money," *Harvard Business Review,* May-June 1977, *55,* 155–161.

Herman, Edward S., and Louis Lowenstein, "The Efficiency Effects of Hostile Takeovers." In Coffee, John, *et al.,* eds., *Knights, Raiders, and Targets.* New York: Oxford University Press, 1987.

Jensen, Michael C., and Richard Ruback, "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics,* April 1983, *11,* 5–50.

Lowenstein, Louis, *What's Wrong With Wall Street: Short-Term Gains and the Absentee Shareholder.* Reading, MA: Addison-Wesley, forthcoming.

Magenheim, Ellen, and Dennis C. Mueller, "On Measuring the Effect of Mergers on Acquiring Firm Shareholders," In Coffee, John, *et al.,* eds., *Knights, Raiders, and Targets.* New York: Oxford University Press, 1987.

Mandelker, Gershon, "Risk and Return: The Case of Merging Firms," *Journal of Financial Economics,* 1974, *1,* 303–335.

Marris, Robin, "A Model of the 'Managerial' Enterprise," *Quarterly Journal of Economics,* May 1963, *77,* 185–209.

Mueller, Dennis C., "A Theory of Conglomerate Mergers," *Quarterly Journal of Economics,* November 1969, *83,* 643–651.

Porter, Michael E, "From Competitive Advantage to Corporate Strategy," *Harvard Business Review,* May-June 1987, *65,* 43–59.

Ravenscraft, David J., and F. M. Scherer, "Divisional Sell-off: A Hazard Function Analysis," working paper, 1987a.

Ravenscraft, David J., and F. M. Scherer, "Life After Takeover," *Journal of Industrial Economics,* September 1987b, *36.*

Ravenscraft, David J., and F. M. Scherer, *Mergers, Sell-offs, and Economic Efficiency.* Washington: Brookings Institution, 1987c.

Ross, David, "Learning To Dominate," *Journal of Industrial Economics,* June 1986, *34,* 337–353.

Scherer, F. M., "Inflation, Capital Budgeting, and the Long View," unpublished manuscript, 1983.

Scherer, F. M., "The Ownership of U.S. Corporations." In *The U.S. Business Corporation: An Institution in Transition,* American Academy of Arts and Sciences, forthcoming.

Shiller, Robert J., "Do Stock Prices Move Too Much To Be Justified by Subsequent Changes in Dividends?" *American Economic Review,* June 1981, *71,* 421–436.

Shiller, Robert, J., "Fashions, Fads and Bubbles in Financial Markets." In Coffee, John, *et al.,* eds., *Knights, Raiders, and Targets.* New York: Oxford University Press, 1987.

Securities and Exchange Commission, Office of the Chief Economist, staff report, "Institutional Ownership, Tender Offers and Long Term Investments," 1985.

Spence, A. M., "The Learning Curve and Competition," *Bell Journal of Economics,* Spring 1981, *12,* 49–70.

Summers, Lawrence H., "Does the Stock Market Rationally Reflect Fundamental Values?" *Journal of Finance,* July 1986, *41,* 591–601.

U.S. House of Representatives, Committee on Energy and Commerce, Subcommittee on Telecommunications and Finance, Hearing, *Impact of Mergers and Acquisitions.* Washington: Government Printing Office, 1987.

Williams, Ronald B., Jr., "Industry Practice in Allocating Capital Resources," *Managerial Planning,* May-June 1970, *18,* 15–22.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 83-103

Monetary and Exchange Rate Policies for International Financial Stability: A Proposal

Ronald I. McKinnon

Without a common monetary standard, the remarkable integration of Western European, North American, and the industrialized Asian economies in both commodity trade and financial flows is less efficient, and becoming untenable. Dissatisfaction with wildly fluctuating relative currency values, euphemistically called "floating" or "flexible" exchange rates, is a prime cause of the resurgence in protectionism.

In 1986-87, the overvalued yen forced Japanese industrialists to close factories, retire workers, and write off once valuable investments in plant and equipment. This parallels what their American counterparts were forced to do between 1981 and 1985, when the dollar suddenly became overvalued. Similarly, the paring down of the British manufacturing base was precipitated when the pound unexpectedly became a strong petrocurrency in 1979-81. In agriculture, Japan and European countries are determined to insulate their domestic prices of farm products from unsettling international influences, including volatile fluctuations in the yen/dollar or mark/dollar exchange rates. Thus, unless exchange stability is first achieved, the American government's attempt to broaden the General Agreement on Tariffs and Trade to encompass agriculture and services is likely to fail.

But what keeps the three major industrial blocs from developing a common monetary standard to prevent exchange-rate fluctuations? Although many people

■ *Ronald I. McKinnon is Professor of Economics, Stanford University, Stanford, California.*

would point to political differences, doctrinal disputes among economists are more important. Well-intentioned politicians and government officials are stymied because of the differing theoretical perspectives of their economic advisers.

The first issue is whether or not a floating foreign exchange market—where governments do not systematically target exchange rates—is "efficient." Because many economists believe that exchange risk can be effectively hedged in forward markets, they argue that international monetary reform is unnecessary.

Second, after a decade and a half of unremitting turbulence in the foreign exchange markets, economists cannot agree on "equilibrium" or desirable official targets for exchange rates if they were to be stabilized. Two separate and contending principles—that of *purchasing power parity* and of *balanced trade*—yield very different estimates for the "correct" yen/dollar and mark/dollar exchange rates.

Third, if the three major blocs can agree to fix nominal exchange rates within narrow bands, by what working rule should the new monetary standard be anchored to prevent worldwide inflation or deflation?

After a brief consideration of some empirical evidence on the magnitude of exchange-rate fluctuations since floating began in the early 1970s, I analyze each of these conceptual issues in the course of demonstrating how the central banks of Japan, the United States, and Germany (representing the continental European bloc) can establish fixed exchange rates and international monetary stability.[1]

The Evidence

In Figure 1, monthly data on the yen/dollar and mark/dollar exchange rates extend from the major breakdown in Bretton Woods parities in 1971 to March 1987. On the vertical scale, percentage changes may be read directly from one month to any other with the zero point representing the mean yen/dollar or mark/dollar exchange rates, 252 and 2.48 respectively, over this sixteen-year period. For example, at its trough of 1.72 marks in late 1979, the dollar was 37 percent below its period mean; at its recent peak of 3.31 marks in February 1985, the dollar was 28 percent above its mean. By late 1987, the dollar had fallen sharply below its 1979 lows.

Casual inspection of Figure 1 indicates that movements in the yen/dollar and mark/dollar exchange rates were not only very large but were also positively correlated. Indeed, the simple coefficient of correlation is .67. In this respect, the unstable dollar is the main problem—suggesting that American monetary policy, in particular, has been insufficiently internationally oriented.

Yet divergent movements between the yen/dollar and mark/dollar exchange rates do occur, as a glance at Figure 1 indicates. In March 1987, for example, the rate of 143 yen/dollar was 57 percent below its period mean, while the rate of 1.81 marks/dollar was "only" 32 percent below its mean. In this sense, in 1985–87 the yen had appreciated substantially against the mark as well as the dollar.

[1] A proposal first broached over a dozen years ago (McKinnon, 1974) in response to the collapse of the Bretton Woods system of fixed exchange-rate parities.

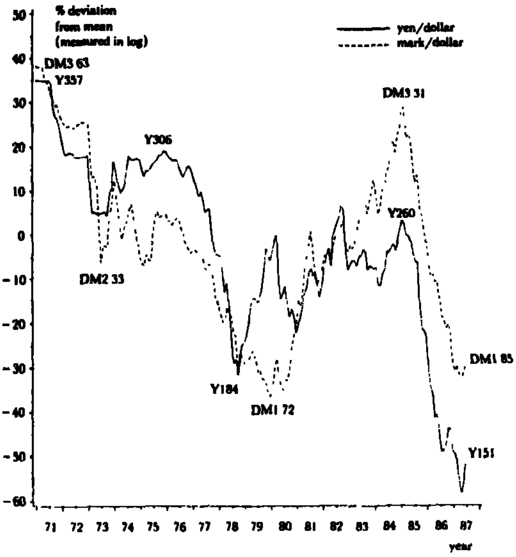


Fig. 1. Mark/dollar and yen/dollar exchange rates: 1971–87 (monthly averages from daily data)
Note: Vertical scale shows percentage changes directly from one month to another. The zero point in the middle of the scale represents the period mean yen/dollar and mark/dollar exchange rates, 252 and 2.48, respectively from 1971 to March 1987.

More generally, also including experiences of smaller "independent" floaters such as the United Kingdom or Australia, what are the stylized facts about currency fluctuations in world markets?

First, relative to profit margins on investment measured in any one national currency, exchange rate changes have been very large: 1 percent in a day, 5 percent in a month, and 20 percent in a year are commonplace (IMF, 1984).

Second, exchange fluctuations have been mainly unanticipated by the market, reflected neither in interest differentials across countries nor in forward premia or discounts in the exchange markets (Frenkel and Mussa, 1980).

Third, these changes have been real in the sense that most domestic prices, including internationally tradable manufacturers, have remained relatively sticky. Because large cyclical fluctuations in exchange rates have not been offset by the much smaller, largely secular, differences in domestic price inflation (Levich, 1986), they induce abrupt changes in international competitiveness.

Fourth, despite the free flow of financial capital, large but variable real interest differentials of 3 to 4 percentage points between similar assets denominated in different currencies are commonplace (Frankel, 1986).

Asset Pricing and Market Efficiency

What causes this great variance in floating exchange rates among industrial economies with open capital markets? The most convincing empirical description is that the exchange rate behaves like a forward-looking asset price: investors continually shift their preferences among yen, mark, dollar, and sterling assets according to how they imagine each exchange rate might move in the future (Frenkel and Mussa, 1980). These expectations are then telescoped back into the spot market to produce the sharp changes shown in Figure 1.

The floating foreign exchange rate seems to be dominated by volatile asset preferences rather than adjusting passively to balance current flows of imports and exports. In the face of uncertainty about the future purchasing power of domestic money, liquid foreign exchange assets are more easily substituted for domestic financial assets (money or bonds) than are physical assets such as real estate or stocks of commodities. Foreign bonds or bank accounts are also convenient hedges against possible future shifts in domestic, political, or commercial risk. These potential capital flows through the foreign exchanges on a daily basis are huge. Since they are so much greater than the value of commodity trade, they dominate observed movements in exchange rates.

When future exchange rates are uncertain, as under a free float, these portfolio preferences become highly sensitive to ongoing "news" about the exchange rate consequences of how, say, future monetary or tax policy in one country might be conducted relative to another. Hence, virtually all the exchange rate variance shown in Figure 1 comes from news that changes expectations—such as the American government's surprise leak to *The New York Times* in mid-January 1987 that it wanted the dollar to depreciate further despite attempts by the Bundesbank and Bank of Japan to support the American currency.

Correspondingly, little or none of the observed variance in the spot exchange rate is predictable from previous knowledge of forward rates (Levich, 1986) or from generally available past information on national money supplies, interest rates, trade balances and so on (Meese and Rogoff, 1983). In fact, the current spot rate seems to embody all relevant news up to that point, and subsequently moves like a random walk driven by new information coming in.[2] Because "excess" profits from speculating on the basis of common knowledge from the past are close to zero, the private foreign exchange market seems to be highly efficient in processing the information available to it.

Nevertheless, I hypothesize that a floating foreign exchange market is *socially inefficient* because private foreign exchange traders face a huge gap in relevant information: the relative future purchasing powers of national fiat monies, none of which has any intrinsic value, are highly uncertain. Thus the assessments of international investors of whether dollar, or yen, or mark assets provide the best combination of yield and safety are unnecessarily volatile.

---

[2] This leaves open the question of whether new information, which shifts current expectations, is accurately borne out in the future. Because exchange rate expectations are so difficult to measure, econometric testing of their accuracy is virtually impossible. Much new "information" about future monetary expansion—as, say, when the president of the central bank suddenly becomes ill—turns out to be ephemeral. Other news, such as changes in the tax laws affecting the relative profitability of holding foreign versus domestic securities, may prove well-founded. Either accurate or inaccurate news, however, can move exchange rates, sometimes dramatically.

The appropriate solution to this portfolio instability would be for the U.S. Federal Reserve System, the Bank of Japan, and the Bundesbank to announce jointly that, henceforth, the three will adhere to a common monetary standard, thereby foregoing the option to inflate at different rates. In effect, they would seek to approximate the conditions prevailing within a single currency area where one monetary unit, say one dollar, would have roughly the same purchasing power over a broad basket of tradable goods in any participating region or country.

To help implement this new policy, fixed nominal targets for the yen/dollar and mark/dollar exchange rates (within narrow bands) would be officially announced, and set to approximate sustainable purchasing power parities. Once achieved, the three central banks would symmetrically adjust their domestic money supplies to maintain these nominal exchange parities and, concomitantly, maintain roughly the same rates of domestic price inflation in internationally tradable goods. Thus speculation on which nation's monetary policy will be the most inflationary becomes pointless; and a major source of instability in international currency preferences is eliminated once speculators know that future nominal exchange rates will be the same as today's.

Although there are many historical precedents, having countries give up some national autonomy over the supply of money in order to secure fixed exchange rates remains controversial. Hence, I shall first show how common arguments favoring exchange rate flexibility become invalid once nations are highly integrated in finance and trade. Then more precise details on what having a "common" monetary policy means, and how it could be facilitated, are spelled out.

"Accidental" Monetary Disruptions from Real Disturbances

Are fixed nominal exchange rates appropriate if "real" shocks affect nations differentially? Most economists still believe that flexibility in nominal exchange rates is necessary to overcome domestic price-level rigidities to balance international payments properly—a belief analyzed in some detail later on. Here, however, consider the reverse problem: how do real shocks impinge on domestic monetary (price-level) stability when economies are highly open on both capital and trade accounts? Is a system of fixed or floating exchange rates the better shock absorber?

Under floating exchange rates, news about future exchange rate movements also encompasses nonmonetary events. For example, a country might reduce its taxation of foreign corporations or the price of its principal export could increase. If the economy of this country is highly open, these events will not warrant a (nominal) appreciation of the country's exchange rate. Instead, if the country's exchange rate is left free to appreciate, such real disturbances would provoke an unexpected (and unwarranted) general price-level deflation in the country. Consider a historical example.

With the worldwide increase in the price of oil in 1979, and Britain's emergence as a major oil exporter, investors projected that U.K. foreign exchange earnings would be greater in the future. Moreover, they also knew that the Bank of England was committed to letting the exchange rate float while following a monetarist rule of limiting the rate of domestic money growth. Because of home currency preference by Britons to convert this future stream of earnings into pounds, sterling's future foreign exchange value was seen to be higher by forward-looking international investors. This was immediately registered in the foreign exchange market by a large increase in the demand for *spot* sterling in 1979. Thus, the nominal (and real) foreign exchange value of sterling rose sharply, precipitately deflating the British economy and reducing international competitiveness so as to wipe out a large chunk of British manufacturing in 1980–81.

In contrast, under a rule of fixed exchange rates, the Bank of England would have stood ready to expand the sterling monetary base and reduce interest rates to satisfy any increased international demand for sterling assets at the agreed-on exchange rate. Moreover, because speculators would know that the Bank was committed to a fixed exchange rate, they would not have anticipated future sterling appreciation in response to the oil price increase. Thus, they would not have so sharply increased their demand for sterling assets in 1979!

The Dutch experience with the discovery and development of large natural gas fields in the early 1960s is an instructive comparison. Because everyone knew that under the old Bretton Woods system the guilder's foreign exchange value was fixed in dollar (and mark) terms throughout the 1960s, no immediate international rush into guilder assets took place—and there was no precipitate exchange appreciation and deflation of the Dutch economy. To be sure, as the natural gas sector expanded and the tax revenues were spent for domestic goods and services thus bidding up wages, other Dutch tradable goods industries inevitably declined—the well-known phenomenon called the "Dutch Disease." But this decline was gradual. The fixed exchange rate avoided sharp appreciation and deflation—accompanied by heavy unemployment—of the kind that Britain was to suffer a decade later.

In summary, agreeing on a common monetary policy resulting in stable nominal exchange rates has two aspects. First, each central bank must give up *discretionary* power to persistently inflate its price level at a rate differing from the common standard on which the group had agreed. Secondly, and less obviously, the cooperating central banks must stand ready to offset the portfolio consequences of internationally differentiated real disturbances to prevent them from *accidentally* disrupting any one country's nominal exchange rate and its price level—as in the British example above and in the American and Japanese cases to be described below.

Only when international investors have such forward assurance will the foreign exchange market be informationally efficient in the social sense. Then private speculation will support the exchange rate targets that the authorities announce.

Incomplete Forward Markets for Goods and Services: The Arrow-Debreu Dilemma

In response to the great turbulence over the past decade and a half, financial markets in Chicago, New York, London, Frankfurt, Tokyo and so on have developed an amazing range of instruments for hedging both exchange and interest rate risk. Now in the late 1980s, capital mobility among the three major blocs is virtually unrestricted, and industrial enterprises may freely borrow or lend foreign currencies directly at virtually any term to maturity. In the words of the Bank for International Settlements (1986, p. 1), "Innovation has improved the efficiency of international financial markets, mainly by offering a broader and more flexible range of instruments both for borrowing and for hedging interest rate and exchange rate exposures. These changes have clearly aided banks and their customers to cope with stresses associated with the greater volatility of exchange and interest rates in recent years."

After glancing at the dazzling array of financial instruments for taking forward exchange positions, most economists might rest assured that most international currency risk associated with trade and investment could be effectively hedged. Nevertheless, firms investing in physical plant or human capital find they have substantial exchange risk which *cannot* be hedged as long as exchange rates are free to fluctuate (Kindleberger, 1972, 1985). Why the paradox?

The problem lies not so much with any inadequacies in the forward markets for foreign exchange, but rather is rooted in the incompleteness of forward contracting in markets for goods and services. We know that a manufacturer contemplating a new investment cannot make all his future production and sales decisions before the plant is built—then lay off the economic risks with a complete set of forward commodity contracts contingent on various uncertain states of nature. Arrow and Debreu (1953; 1959) have taught us the critical importance of forward commodity markets to hedge against price risk, coupled with insurance markets to offset production risk.

However, a dilemma arises because the transactions costs and moral hazard associated with contract enforcement of such contingent futures are generally prohibitive. Usually, producers cannot feasibly protect themselves with outside insurance policies against their inability to deliver, nor against unexpected changes in cost (Arrow, 1973; Greenwald and Stiglitz, 1986). A farmer would not necessarily want to sell his crop forward years—or even months—in advance if he cannot predict his own output nor obtain crop insurance against a multiplicity of possible natural disasters (McKinnon, 1967).

Thus producers typically sell forward only some of their output—and that but a few weeks or months in advance on a noncontingent basis. Many goods are simply held in inventory until customers come in to buy spot.[3] Because of these incomplete forward markets for commodities, therefore, fixed investments for producing internationally tradable goods and services cannot be hedged effectively against most foreign exchange risk.

Let us illustrate this important point with the example of an industrialist contemplating where to build a new plant to produce laser disks. As in a common currency area, he must identify that geographic locale where (domestic currency) factor costs are relatively low compared to projected output prices—what I shall call the "economic fundamentals." But with floating exchange rates, he must also give positive weight to that country whose currency is relatively undervalued by the purchasing power parity criterion: the "real exchange rate" effect. Regardless of how our laser disk manufacturer balances the "economic fundamentals" with the "real exchange rate" to select a country for production, the choice can easily be made obsolete by currency fluctuation.

---

[3] On the buyer's or consumer's side, a symmetrical inability to protect against future uncertain events—say the appearance of new and better products—limits their willingness to buy forward. Incomplete contingent forward markets pushes them toward buying spot.

Only if the planned production of laser disks could be insured *and* sold forward over the lifetime of his plant could our industrialist absolve himself of exchange risk. Suppose these hypothetical forward prices for laser disks over a continuum of future dates were specified in all other relevant currencies. Then, the existing forward markets in foreign exchanges could be fully utilized. The industrialist could sell (go short) other currencies for the currency in the country of production at the various terms to maturity and amounts negotiated in any forward commodity contracts. Because of the Arrow-Debreu dilemma, however, such forward commodity contracts, fully insured against unexpected dislocations and delays in production, do not exist.

To be sure, a merchant exporter with inventories of already-produced goods can hedge against short-term exchange risk when he sells consignments abroad to be delivered in, say, 30 days or 6 months. After forward selling his goods for foreign money, he can effectively "double hedge" by selling foreign exchange forward to get safely back into the domestic currency (McKinnon, 1979; Kawai and Zilcha, 1986). And this type of forward currency transacting is very important in ameliorating short-term exchange risk.

But short-term hedging on a noncontingent basis covers only a small proportion of the potential longer term exchange risk arising from investment decisions that are made today. In effect, incomplete forward commodity markets exposes industrialists, with long-term fixed investments to foreign exchange risk which they cannot avoid. Thus, random variance in exchange rates introduces new net risk into the world economy. In addition to the British overvaluation of 1979–81, consider more recent examples of this fundamental unhedgability of long-term investment commitments against foreign exchange risk. When the dollar was generally weak in the 1970s, and became substantially undervalued from 1977 to 1980, American tradable good industries looked profitable and "excessive" investments occurred in certain kinds of U.S. mining and manufacturing, with agriculture also becoming overcapitalized. As the dollar (unexpectedly) rose in 1981 and became overvalued until late 1985, these industries then suffered a big shakeout with bankruptcies and plant closures. Obviously, these companies did not (could not) hedge their foreign exchange positions.[4]

The resulting avalanche of protectionist sentiment in the U.S. Congress is still with us. Although by 1986 and into 1987 the dollar is no longer overvalued against the yen and continental European bloc of currencies, industrialists see considerable future exchange-rate uncertainty in assessing whether or not to rebuild America's productive capacity in tradable goods industries. Investment is inhibited.

---

[4] Note that even companies which do not themselves engage in foreign trade can be highly vulnerable to such exchange rate fluctuations (Hodder, 1983).

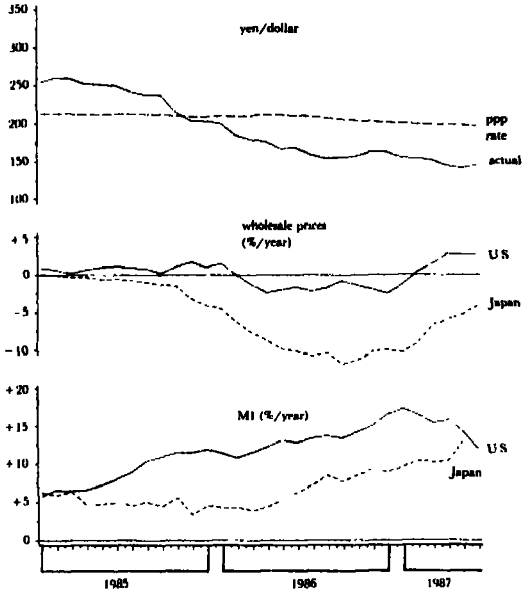


Fig. 2. Exchange rates, prices and money: Japan and the U.S. 1985–87
Note: Rates of change in money supplies and wholesale prices are year over year estimates taken from *The Economist*; estimates of Purchasing Power Parity Exchange Rates (PPP) from Ohno, 1987 (projected into 1987 from Japanese and American wholesale price indexes).

Similarly, as the yen rose incredibly from 260 yen/dollar in March 1985 to around 140 yen/dollar in the early part of 1987, Japanese industrial output turned down and much of her previously installed manufacturing capacity suddenly became unprofitable. In 1986, there occurred an unprecedented deflation of over 10 percent in the yen prices of Japanese industrial goods as measured by Japan's official wholesale price index (Figure 2). The resulting industrial depression in Japan has prompted Japan's preeminent industrialist, Mr. Akio Morita (1986), President and Chairman of the Sony Corporation, to call for reforms such that national money becomes "a common scale of value internationally rather than just another speculative commodity." Otherwise, he cannot properly decide in which goods to invest, in which country to produce them, or how to arrange for future sales and supplies.

To be sure, neither American industrialists in 1977–80, nor Japanese industrialists prior to 1986–87, necessarily made wrong investment choices ex ante. Rather, they faced incomplete information regarding the future course of their own country's exchange rate—an information gap that turned out to be devastating.

A Fleeting Vision of How Monetary Cooperation Could Work

When the dollar soared to 260 yen and 3.3 marks in late February 1985, the unsustainable loss in America's international competitiveness since 1981 had become so great as to be obvious to everybody. The Reagan administration finally abandoned

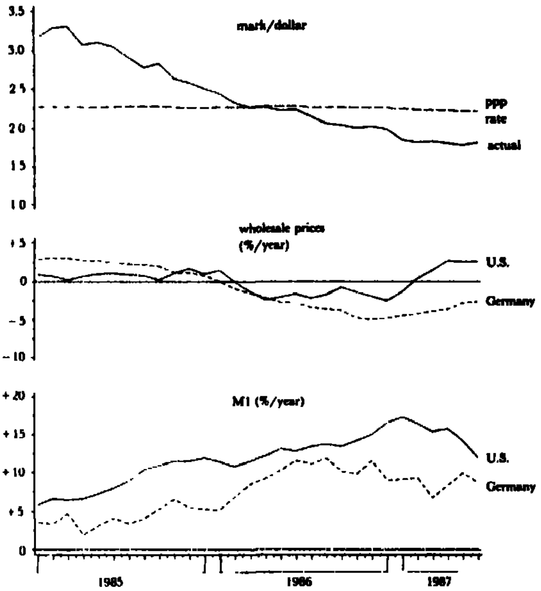


Fig. 3. Exchange rates, prices and money: Germany and the U.S. 1985–87

Note: Rates of change in money supplies and wholesale prices are year over year estimates taken from *The Economist*; estimates of Purchasing Power Parity Exchange Rates (PPP) from Ohno, 1987 (projected into 1987 from German and American wholesale price indexes).

its longstanding refusal to treat the dollar exchange rate as a legitimate target for public policy.

For the rest of 1985, central banks cooperated to nudge the dollar down directly in the foreign-exchange markets—a policy that was announced and ratified by the Plaza Hotel Accord in September 1985. More important, their domestic money growth rates supported this objective. Figures 2 and 3 show that growth in M_1, the basic money supply, rose sharply in the U.S. to more than 10 percent per year, while Germany and Japan followed relatively tight money policies with M_1 growth less than 5 percent.

At 200 yen and 2.3 marks by the end of 1985, the dollar was correctly aligned with the currencies of America's Japanese and European trading partners in two closely related respects. First, there was approximate purchasing-power parity. One dollar could buy at the wholesale level a broad basket of manufactured goods in Japan and Europe similar to what it could buy at home (Ohno, 1987).

Second, price inflation in tradable goods (as measured by changes in wholesale-price indexes) were virtually the same in the three areas at the end of 1985, and close to zero. Being properly aligned, the yen/dollar and mark/dollar exchange rates were not themselves imposing any significant upward or downward presure on the average price level of tradable goods in any of the three major blocs (Figures 2 and 3).

Hence, late 1985 into early 1986 presented a wonderful opportunity for the finance ministers of the Group of Seven industrialized democracies to reinstitute a common monetary standard. At that time, a formally announced policy whereby the dollar would, henceforth, be kept within fixed narrow bands of, say, 190 to 210 yen and 2.2 to 2.4 marks, would have been easy to implement. If the dollar fell toward the lower intervention points, The U.S. Federal Reserve System would be obligated to tighten American monetary policy just as the Bank of Japan and Bundesbank (representing the European bloc) would be obligated to ease, and vice versa should the dollar become unduly strong.

To keep potentially volatile exchange rates within their prescribed bands, the three central banks must control *relative* short-term interest rates on a weekly basis, or even a daily one. Through open market operations (or discounting), the weak currency country would raise short-term interest rates to attract capital from abroad, and vice versa for the strong currency countries. With confidence in the official parities, very small changes in interest rates would attract (or repel) sufficient short-term capital to equilibrate the foreign exchanges.

Only as a last resort, or because of unusual financial turbulence, would substantial direct intervention in the foreign exchanges be necessary. Then, to be decisive, *symmetrically unsterilized intervention* is appropriate (McKinnon, 1984). For example, if the Bank of Japan intervened at the lower bound for the dollar to purchase dollars and sell yen, the American monetary base would contract as the Japanese base expanded—leading to further equilibrating changes in relative interest rates in both countries.

Once speculators understood these new rules of the game, containing unambiguous information regarding the mutual intentions of the major central banks, then private short-term capital flows would help to stabilize exchange rates within their official bands. Very little actual monetary adjustment on a daily or even weekly basis would be necessary as long as everyone knew that decisive official action would be forthcoming if the need arose.

In the longer run, the three central banks would aim to keep constant their common (wholesale) price level in internationally tradable goods—whether measured in yen, dollars, or marks—as the nominal anchor for the system as a whole. For this purpose, aggregate money growth, the sum of domestic credit expansions, could be targeted as an intermediate variable (McKinnon, 1982; 1984). For example, if international deflation threatened, then joint money growth would be slowly accelerated.[5]

Unfortunately, this naive vision was not realized. Instead of being stabilized at its purchasing-power-parity level at the end of 1985, the dollar continued to fall throughout 1986 and then plunged below 140 yen and 1.8 marks in early 1987. Instead of feeding the foreign-exchange markets clear information regarding their mutual monetary intentions, governments criticized each other's alleged policy failure and then appeared to act at cross purposes. Despite the weak dollar, growth in the American money supply throughout 1986 remained far in excess of that in Germany and Japan, as shown in Figures 2 and 3.

---

[5] Note that interest rates are much less useful as monetary indicators for the longer run problem of securing price level stability for the group. Because of unstable inflationary expectations, sometimes called Fisher effects, absolute levels of nominal interest rates are ambiguous for measuring whether money is too "tight" or too "easy." Hence the need to pay attention to the aggregate money growth of the triumvirate to help anchor the common price level in the longer run, even while control over relative short-term interest rates is assigned to stabilizing exchange rates.

The Exchange Rate and the Trade Balance

What provoked the American government to push and to keep pushing the dollar down so far—thus imposing undue deflationary pressure on Japan and continental Europe while risking the resurgence of inflation in the United States? The pursuit of what is now a false academic doctrine: that a once-and-for-all devaluation of a currency can systematically reduce that country's trade deficit as per the old Robinson (1937) and Meade (1951) elasticities approach to the balance of trade. When industrial economies were still fairly insulated from one another in the late 1930s into the 1950s, one could reasonably presume that a devaluation would improve the trade balance. But among highly open economies of the 1970s and 1980s, exchange-rate changes have no predictable effect on the money value of net trade balances because of their more pervasive macroeconomic repercussions (Miles, 1979; McKinnon, 1981; McKinnon and Ohno, 1986).

After the sharp appreciations of the mark and yen over the past two years, most economists expected that the German and Japanese loss of "international competitiveness"—that is, relatively higher prices of their goods on world markets—would reduce their trade surpluses as measured in dollars. These direct price effects did not have a strong enough effect on the flow of imports and exports for several reasons.

First, their improved terms of trade meant that the Japanese and Europeans earned more dollars per unit of exports without paying any less per unit of imports.

Second, there were internal deflations in Japan and Germany (Figures 2 and 3) that were a necessary consequence of their exchange appreciations. These deflations so depressed Japanese and German demand for imports and released more goods for export that the physical volume of exports relative to imports did not fall sufficiently (although substantial real adjustment did occur) to offset their more favorable terms of trade. Consequently, both the Japanese and German net trade surpluses measured in dollars actually increased in 1986 and 1987 in comparison to 1985.

Granted, the current huge excess of imports over exports in the American economy is unsustainable and must eventually be corrected one way or another. But this trade deficit of $150 billion to $200 billion a year merely reflects the saving-investment gap in the American economy created by the not coincidentally equally large U.S fiscal deficit. Given the still-commanding reserve-currency status of the U.S. dollar in world finance, the American deficit acts as a huge vacuum cleaner that sucks up other countries' savings through the preemptive issue of the U.S. Treasury bonds in world financial markets. Consequential macroeconomic adjustments then force them to generate dollar surpluses in their commodity trade, whether or not they have appreciated their exchange rates. Getting rid of this chaos in the American public finances is the only satisfactory way to reduce the trade deficit. Because the great 1985–87 depreciation of the dollar has had, and will have, no substantial impact on this structural fiscal imbalance and consequent saving shortage in the American economy, it is incapable of correcting the U.S. trade deficit short of precipitating a major collapse of domestic investment within the United States.[6]

To be sure, the trade deficit is an important subject worthy of intense negotiation —but only in the context of assigning fiscal policies to resolve it. And in a world where real interest rates are still too high, there is little doubt that the United States should sharply cut its fiscal deficit to reduce its borrowing, rather than trying to persuade other governments to expand their expenditures and cut tax revenues.

For the international system to hold together, the U.S. government must give up its false, although widely held, view that dollar devaluation will itself substantially improve the dollar value of the net trade balance. For industrial countries, economists should jettison the whole elasticity approach relating the exchange rate to the balance of trade. Not only is this obsolete doctrine currently distorting relative currency values, but it suggests a "need" for continual exchange rate flexibility—in order to balance ever-changing flows of imports and exports—thus making a common monetary standard impossible.

Purchasing Power Parity and the Nominal Anchor

But financial markets can be thrown into turmoil if participants believe that policy makers have half a mind to follow an incorrect theory and drive the dollar down. The foreign exchange crisis from January through May 1987 is a case in point. Because the markets came to believe that American officials wanted the dollar to depreciate indefinitely, the normal inflow of private capital to cover the U.S. trade deficit dried up, thus necessitating massive official interventions by central banks to support the dollar. At the same time, long-term interest rates in American bond and mortgage markets rose sharply.

To get out of this unfortunate syndrome, we need an alternative theory of how exchange rates should be established. Purchasing power parity (PPP) can be an unambiguous theoretical guide for central banks and one with which the private financial markets can also feel comfortable.

The central idea behind PPP is to calculate nominal exchange rates that would align national price level of internationally tradable goods as approximately measured

---

[6] This ambiguous impact of an exchange rate change on the trade balance is not just a short-run consequence of sticky invoice prices for imports and exports—the familiar J-curve effect. In principle, adverse expenditure effects could offset direct price effects for several years (McKinnon 1981). In the very long run, of course, a devaluation simply washes out as internal price levels adjust and purchasing power parity is restored.

Table 1
Exchange rates, PPP rates, and target ranges

	Actual exchange rates (quarterly averages)		*Purchasing power parity rates*	
	yen / $	*mark / $*	*yen / $*	*mark / $*
1986 IV	160	2.01	205	2.24
1987 I	153	1.84	202	2.22
I	143	1.81	199	2.18
III	147	1.84	201	2.18
IV	136^{P}	1.68^{P}	200^{P}	2.17^{P}
	Suggested target ranges			
1988 I	160 to 180	1.9 to 2.1	199^{P}	2.16^{P}

Note: P indicates projected.
Source: OECD producer prices, IMF exchange rates PPP calculations from McKinnon and Ohno, 1987.

by their respective producer or wholesale price indexes. Table 1 shows that concurrent purchasing power exchange rates in late 1987 are about 200 yen/dollar and 2.18 marks/dollar.[7] If the internal paths for price inflation were roughly the same (as they had been at the beginning of 1986) in Germany, Japan, and the United States, then these exchange rates (within narrow bands) should be the officially announced targets.

In 1986 and 1987, however, the prolonged undervaluation of the dollar—i.e., overvaluation of the yen and, to a lesser extent, the mark—has put the U.S. on a higher inflation path (Figures 2 and 3). In Japan, deflationary pressure from the overvalued yen has caused a decline in a broad index of Japanese producer (and wholesale) prices and depressed the normal (high) rate of growth in Japanese money wages. Indeed, American money wages are now increasing one to two percentage points faster, despite Japan's higher trend in labor productivity.

However, the momentum in these wage and price movements can only be altered with a lag. Thus, to facilitate smoother convergence to a common price level, exchange-rate targets should be set somewhat below current PPP levels, say in the neighborhood of 170 yen/dollar and 2.0 marks/dollar in late 1987. After the dollar had been nudged upward by relative monetary adjustments, unchanging official bands of, say, 160 to 180 yen/dollar and 1.9 to 2.1 marks/dollar could be established for early 1988 (see Table 1). In 1988, if the exchange rate misalignment with its differential inflation rates were to persist, then these target bands should be lowered somewhat.

---

[7] Without directly comparable national price indexes, calculating PPP exchange rates for internationally tradable goods can only be indirect and approximate. Nevertheless, new methods have been evolving (Ohno, 1987; McKinnon and Ohno, 1987) based on either WPIs or PPIs that, for the 1980s, seem to be significantly more efficient than the traditional technique based on the presumption that purchasing power parity held in some arbitrarily chosen base year.

Then, with exchange rates known to be fixed into the indefinite future, international commodity arbitrage and mutual monetary adjustment would ensure convergence to the same rate of commodity price inflation (preferably zero) in all three countries. Tradable goods prices (PPIs) would then be aligned close to purchasing power parity and relative growth in national money wage claims would eventually reflect differentials in productivity growth, as had been the case in the 1960s under the old Bretton Woods System of pegged exchange rates when Japanese money wages grew much faster than those in the United States or Germany.

However, this proposal differs from the Bretton Woods agreement, and also from subsequent American behavior, because the Fed would actively respond to the strength or weakness of the dollar against marks and yen, rather than simply waiting for the Bundesbank and the Bank of Japan to adjust. Although this will on occasion upset political sensibilities over short-term rates, it will pay political dividends in more stable long-term bond and mortgage yields once investors know that the dollar exchange rate will be roughly the same next year, or 10 years from now.

For example, for late 1987 into 1988, raising the yen/dollar and mark/dollar exchange rates toward their PPPs requires the U.S. Federal Reserve System to follow a relatively tight money policy while the Bank of Japan and the Bundesbank are more expansionary. As the mark and yen fall toward their PPP levels, inflationary pressure on prices and wages in the United States should slacken while it picks up in Japan and Europe. But the nominal anchor, the common price level measured at PPP exchange rates, would not be upset if this monetary adjustment were indeed symmetrical.

Announcement effects are important in order for private expectations to coalesce around the new exchange rate targets. If the Fed simply tightens without revealing its foreign exchange objectives, more severe monetary contraction and higher short-term interest rates might be necessary to boost the dollar toward PPP. Because expectations effects would be much less favorable, American long-term interest rates are less likely to come down as quickly.

Once exchange rates are properly aligned (according to PPP), representatives of the Fed, the Bank of Japan, and the EMS bloc should meet continually to monitor the behavior of their common price level in internationally tradable goods (as measured by PPIs or WPIs). Fortunately, information on commodity prices, even producer prices, becomes available much faster than the GNP statistics so widely watched today. Collective monetary expansion would be tailored downward or upward according to inflationary or deflationary trends in the common price level.

For example, the international integration of financial markets implies that stock market crashes, as occurred in October 1987, are felt in concert by all the industrial countries. In the face of such a deflationary threat, collective monetary expansion is warranted—without allowing any one country to get its currency undervalued and secure a mercantilistic advantage over its neighbors.

Even better as a nominal anchor for the system as a whole, a broadly based international price index, similar in scope to a PPI, could be priced out in each currency using a common set of representative weights for the industrial economies. A subcomponent consisting of homogeneous primary products, which are priced to market each day, could serve as an early warning indicator (Heller, 1987) for movements in the broader index.

Participating central banks could use this nominal anchor as their own internal price-level target. For example, the Bundesbank could use the mark value of an international PPI as its target for "zero" domestic price inflation. If the Fed and Bank of Japan adopted similar internal targets based on the dollar and yen values of an international PPI, consistency between domestic price-level objectives and the obligation to maintain fixed exchange rates would be assured. (In the meantime when no such international index is available, existing national PPIs can be used to approximate when the common price level is rising or falling (McKinnon and Ohno, 1987).)

Conversely, to prevent unnecessary conflict, countries should not adopt internal nominal anchors—such as the Consumer Price Index or GNP deflator (both containing nontradable services) or growth in nominal GNP itself—that may be inconsistent with their obligation to maintain fixed nominal exchange rates.

The Transfer Problem

Under fixed exchange rates, international commodity arbitrage and coordinated monetary adjustments would always maintain a rough alignment in tradable goods prices across the three countries. Nevertheless, depending on relative national imbalances between saving and investment, trade deficits or surpluses would develop continually. For example, American dissaving could lead to a transfer of capital from Japan to the United States. How would real adjustment in the trade balance, in response to an international capital flow, be effected once fixed exchange rates were in place?

It is important to distinguish two kinds of relative price or "real exchange" rate changes that are often confused in the literature: (1) the terms of trade or the value of a broad basket of tradable goods of Country A vis-à-vis that of Country B; and (2) the price ratio of tradable goods to nontradable goods and services *within* a country. Among diversified industrial countries, only modest changes of the second kind—that is, the ratio of tradable to nontradable prices—are relevant for effecting the net transfer of capital from one country to another (Kim, 1987).

Suppose that under fixed exchange rates, American dissaving led to a capital inflow from Japan. Then the increased American absorption—that is, domestic consumption and investment expenditures—from domestic credit expansion would cause some (slow) bidding up of nontradable prices (largely services) in the United States. Similarly, the fall in credit expansion in Japan would induce a decline in Japanese absorption and a slower rate of increase in Japanese nontradables prices relative to tradables. Although these relative price movements within both countries would be modest, gradual, and need not be permanent, they would be sufficient to support the transfer of saving from one highly open economy to another, as with past experiences of fixed exchange-rate regimes approximating a common currency area.

However, the new monetary standard would rule out precipitate changes in any one country's terms of trade from exchange rate fluctuations arising out of capital flows from one country to another, such as the sharp appreciation of the dollar in the early 1980s. Such "false" changes in the terms of trade can actually impede the transfer process. By depressing absorption in those countries with appreciated currencies and stimulating expenditures in those countries who get their currencies undervalued, absorption adjusts the "wrong way" in response to a capital inflow, possibly delaying the emergence of a trade deficit in the capital importing country.

Bretton Woods Revisited

What does history tell us about industrial economies voluntarily agreeing to fix exchange rates? The relevant examples are the late 19th-century gold standard from 1879 to its sudden demise in August 1914, the Bretton Woods system from 1945 to 1971, and the European Monetary System (EMS) from 1979 to the present. All three were successful in sustaining a high level of international trade, while avoiding major changes in real exchange rates, like those of the last decade, that so undermine investment efficiency and aggravate protectionism. Yet no single historical example is a complete "model" for our current most important problem: to fix the mark/dollar and yen/dollar exchange rates while anchoring their common price level.

True, the Bretton Woods system succeeded in virtually fixing exchange rates among Germany, Japan, and the United States from 1950 to 1970. But this result was more by accident than by design because the 1945 agreement favored some (controlled) exchange flexibility and national monetary independence.

The Bretton Woods authors—principally Britain's John Maynard Keynes and America's Harry Dexter White—imagined that exchange rates should be adjusted from time to time to correct trade imbalances and/or to accommodate differential rates of national price inflation. Under the prevailing Keynesian theory, being able to choose your own rate of price inflation was then thought to be useful in preventing unemployment and targeting investment at the national level. Of course, that view is no longer so much in vogue among professional economists.

But Keynes and White also wanted to prevent unwarranted changes that might give one nation a mercantilistic advantage. Hence the seeming contradiction in the official interpretation of the Bretton Woods Agreements that "par values for exchange rates should be stable but adjustable."

In the circumstances from the late 1930s into the early 1950s, occasional controlled changes in par values were feasible and potentially effective in helping to balance international payments because trade was limited and capital flows were restricted (McKinnon and Ohno, 1986). Thus governments should negotiate about "necessary" changes in par values without rates changing in an uncontrollable manner beforehand with huge losses in official exchange reserves.

Similarly, within the EMS of the 1980s, occasional small changes in exchange parities have effectively offset price inflation that has been higher in France and Italy than in Germany. Because exchange controls still partly insulate the French and Italian capital markets, these small changes are still feasible, although they have become increasingly traumatic as European financial markets become more integrated. (Nevertheless, by committing themselves in advance to a system of par values for exchange rates, the EMS has been remarkably successful in promoting a convergence of monetary policies toward a common standard (Ungerer *et al.*, 1986). The previously high inflation rates in France and Italy have come down towards the much lower level prevailing in West Germany.)

In contrast, the German, Japanese, and American financial markets are completely open in the 1980s. Hence any negotiations over obvious changes in par values would provoke massive speculative capital flows. Indeed, fixed exchange parities broke down in 1971 precisely because the United States insisted on negotiating a devaluation to inflate faster than her European and Japanese allies, a "right" which was consistent with the economic philosophy underlying the 1945 Bretton Woods articles.

Hence, in foregoing the right of each participating nation to inflate at different rates, the operative economic philosophy behind my fixed exchange rate proposal would be closer to that of a gold standard than to Bretton Woods.

Other economists have suggested different approaches to managing exchange rates. The interesting proposals for flexible target zones for exchange rates put forward by John Williamson (1985) are much more in the spirit of Bretton Woods. Williamson sees controlled exchange rate changes as sometimes being necessary to offset unforeseen changes in a country's terms of trade or other real shocks. In effect, Williamson wants to keep open the option for occasional official adjustments in par values. Hence, his system would remain vulnerable to speculative attack as international investors tried to guess how exchange rates in the future might differ from those now prevailing. Nominal interest rates, particularly at long term, would become very sensitive to anticipations of any official changes in exchange rate parities, even when they were gradual as with a crawling peg. More fundamentally, for investment decisions that are made today, the national money would lose its quality of being a stable long-run standard of international value if (official) exchange rates changed in the future.

In a logically consistent but not empirically practical vein, a second group of economists—Tobin (1982), Dornbusch (1983) and Bergsten (1983)— have toyed with the idea of reimposing exchange controls on the capital account to gain greater stability in exchange rates without giving up exchange flexibility in the long run. By restricting the operation of the international capital market to support monetary nationalism, they want to make the world look more like that seen by Keynes and White in the late 1930s.

An International Gold Standard Without Gold

What was the essential strength of the 19th century gold standard? Unlike other commodities, did the yellow metal have some intrinsic property for keeping a constant purchasing power over other goods and services through time? Did the requirement of gold convertibility prevent badly motivated governments from (mis)managing the issue of national currency?

None of these common arguments in favor of a return to the gold standard carry much weight or fineness. In the 19th century, new gold discoveries tended to create bursts of worldwide inflation—albeit modest by modern criteria; then, the adoption of, or return to, the gold standard by the United States, Germany, Japan, and other countries in the 1870s and 1880s created a shortage of gold reserves and unwarranted worldwide deflation(s) (Cooper, 1982). Throughout the 19th century, greater deflationary pressure was avoided only because paper and deposit monies were substituted for gold coin on a massive scale (Triffin, 1964).

Instead, the late 19th century monetary system worked reasonably well, despite considerable instability in the underlying gold base, because of its *international* character. To defend its national gold parity, each nation came to regulate domestic money issue according to its balance on international payments. Surplus countries automatically expanded their money supplies while those in deficit contracted. Almost by accident, this resulted in generally fixed nominal exchange rates within narrow bands called the gold points for each pair of national currencies, in purchasing power parity, and in similar rates of domestic price inflation in each country.

This common monetary standard set the stage for the great expansion in trade among industrial and primary products producing countries, and for vast transfers of European capital through the London capital market to North America, Latin America, Australasia and so forth without exchange rate changes and without misaligning national price levels. Because speculators know that national money supplies would be adjusted to support the international standard, stabilizing capital flows served to keep exchange rates within their gold points with very little official intervention being required.

Without actually returning to a gold standard, such monetary cooperation resulting in fixed parities and roughly the same underlying inflation rate is quite feasible in the late 1980s. The positive international aspects of the late 19th century system can be reestablished without the disadvantage of an unstable monetary base, where gold coexists uneasily with bank notes, bank deposits, money market mutual funds, and so on. In the proposed new regime as described, the aggregate monetary base of the three blocs would be consciously managed to prevent worldwide inflations or deflations.

Indeed, achieving price stability cooperatively under fixed exchange rates is likely to be easier to manage than if each central bank acts unilaterally under floating rates. Because private portfolio preferences are now so volatile across dollar, yen, and mark assets, any one central bank cannot easily estimate the demand for its own money and then do a good job of stabilizing its own price level. Unexpected exchange rate changes can impose either more inflation or deflation than previously anticipated. When exchange rates and portfolio preferences across the major currencies are stabilized, however, the impact of joint monetary expansion or contraction on the common price level could well become more predictable.

Once America, Japan, and Europe negotiate such an agreement and their management of the joint monetary base is seen to be successful, most smaller countries (which are not prone to chronic inflation) would voluntarily decide to fix their own exchange rates to the dollar, mark, or yen. In effect, the unique American role (burden?) in the 1950s and 1960s of providing a reserve currency for the rest of the world would now be shared with the EMS group and Japan.

■ *I would like to thank Stanley Fischer, Jacob Frenkel, Marc Miles, Kenichi Ohno, Jurgen Schroder, and Joseph Stiglitz for their helpful comments clarifying various parts of the manuscript. However, they are neither implicated in, nor necessarily committed to, the proposal itself.*

References

Arrow, Kenneth, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Review of Economic Studies*, 1964, *XXXI*, 91–96. (Translated from *Econometrie*, 1953).

Arrow, Kenneth, *Essays in the Theory of Risk Bearing*. Chicago: Markham Press, 1973.

Bank for International Settlements, *Recent Innovations in International Banking*, Group of Ten Industrial Countries, April 1986.

Bergsten, C. Fred, "Currency Crisis," *The New York Times*, September 23, 1983.

Cooper, Richard, "The Gold Standard: Historical Facts and Future Prospects," *Brookings Papers on Economic Activity* I, 1982.

Debreu, Gerard, *Theory of Value*. New York: John Wiley and Sons, 1959.

Dornbusch, Rudiger, "Flexible Exchange Rates and Interdependence," International Monetary Fund, *Staff Papers*, March 1983, 3–38.

Frankel, Jeffrey, "International Capital Mobility and Crowding Out in the U.S. Economy." In Hafer R., ed., *The Increasing Openness of the U.S. Economy*. Federal Reserve Bank of St. Louis, 1986, Ch. 2, 33–69.

Frenkel, Jacob, and Michael L. Mussa, "The Efficiency of Foreign Exchange Markets and Measures of Turbulence," *American Economic Review*, May 1980, *70*, 374–81.

Greenwald, Bruce, and Joseph Stiglitz, "Externalities in Economies with Imperfect Information and Incomplete Markets," *Quarterly Journal of Economics*, May 1986, 229–264.

Heller, H. Robert, "Anchoring the International Monetary System," unpublished paper, Board of Governors of the Federal Reserve System. Washington, DC, March, 1987.

Hodder, J. E., "Exposure to Exchange-Rate Movements," *Journal of International Economics*, November 1982, *13*, 375–386.

International Monetary Fund, "Exchange Rate Volatility and World Trade," *Occasional Paper* 28, July 1984.

Kawai, Masahiro, and Itzhak Zilcha, "International Trade with Forward-Futures Markets Under Exchange-Rate and Price Uncertainty," *Journal of International Economics*, February 1986, *20*, 83–98.

Kim, Yoonbai, "International Transfers of

Capital and the Role of the Terms of Trade," unpublished paper, Economics Department, Stanford University, February 1987.

Kindleberger, Charles P., "The Benefit of International Money," *Journal of International Economics*, September 1972, 425–42.

Kindleberger, Charles P., "The Dollar Yesterday, Today and Tomorrow," *Quarterly Review*, Banca Nazionale del Lavoro, December 1985, *2*, 295–308.

Levich, Richard, "Gauging the Evidence on Recent Movements in the Value of the Dollar." In *The U.S. Dollar: Recent Development, Outlook, and Policy Options*, Federal Reserve Bank of Kansas City, 1986, Ch. 1, 1–28.

McKinnon, Ronald I., "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, *72*, 320–33.

McKinnon, Ronald I., "The Exchange Rate and Macroeconomic Policy: Changing Postwar Perceptions," *Journal of Economic Literature*, June 1981, *XIX*, 531–57.

McKinnon, Ronald I., "Futures Markets, Buffer Stocks, and Income Stability for Primary Producers," *Journal of Political Economy*, December 1967, *75*, 844–61.

McKinnon, Ronald I., *An International Standard for Monetary Stabilization*. Washington, DC: Institute for International Economics, 1984.

McKinnon, Ronald I., *Money in International Exchange: The Convertible Currency System*. New York: Oxford University Press, 1979.

McKinnon, Ronald I., "A New Triparite Monetary Agreement or a Limping Dollar Standard?" *Essays in International Finance*, Princeton University, October 1974, *106*.

McKinnon, Ronald I., and Kenichi Ohno, "Getting the Exchange Rate Right: Insular Versus Open Economies," unpublished paper, Economics Department, Stanford University, December 1986.

McKinnon, Ronald I., and Kenichi Ohno, "Purchasing Power Parity as a Monetary Standard" unpublished paper, Economics Department, Stanford University, September 1987.

Meade, James E., *The Balance of Payments*. London: Oxford University Press, 1951.

Meese, Richard, and Kenneth Rogoff, "Empirical Exchange Rate Models of the 1970s: Do They Fit Out of Sample," *Journal of International Economics*, February 1983, *14*, 3–24.

Miles, Marc A., "The Effects of Devaluation on the Trade Balance and the Balance of Payments: Some New Results," *Journal of Political Economy*, June 1979, *87*, 600–620.

Morita, Akio, presentation to *International Working Round on Exchange Rates and Coordination*, Zurich, Switzerland, June 19, 1986.

Ohno, Kenichi, "Estimating Purchasing Power Parities in the 1970s and 80s: The Price Pressure Approach," unpublished paper, Economics Department, Stanford University, September 1987.

Robinson, Joan, "The Foreign Exchanges." In *Essays in the Theory of Employment*. New York: MacMillan, 1937, Ch. 1.

Tobin, James, "The State of Exchange Rate Theory: Some Skeptical Observations." In Cooper, Richard, *et al.*, *The International Monetary System Under Flexible Exchange Rates*. Cambridge: Ballinger, 1982, 115–28.

Triffin, Robert, "The Evolution of the International Monetary System: Historical Reappraisel and Future Perspectives," Princeton *Studies in International Finance* No. 18, June 1964.

Ungerer, Horst, *et al.*, "The European Monetary System: Recent Developments," *Occasional Paper* No. 48, International Monetary Fund, December 1986.

Williamson, John, "The Crawling Peg," *Essays in International Finance*. Princeton University, 1965, No. 50.

Williamson, John, *The Exchange Rate System*. Revised. Washington, D.C.: Institute for International Economics, 1985.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—pages 105–112

Doubts About the McKinnon Standard

Rudiger Dornbusch

For over ten years, Ronald McKinnon has advocated a new monetary system centered on fixed exchange rates between the Japanese yen, the German mark and the U.S. dollar. Unperturbed by the collapse of the fixed exchange rate system in 1971–73, McKinnon was already arguing in 1974 for new monetary arrangements based on rigidly fixed rates but built on stronger foundations—world monetarism rather than national monetary discretion (McKinnon, 1974). Since then he has refined his plan for monetary integration under rigidly fixed exchange rates—"a gold standard without gold" as he recently called it (McKinnon, 1987b).

His proposal has fallen on fertile soil because the dramatic volatility in exchange rates makes the financial press and the business community grasp for ready answers. McKinnon's scheme seems to solve the problem of currency instability and to judge from his presentation there are only merits and not one shortcoming. In fact, however, there is little theoretical or empirical basis for his standard. To support this criticism, I will summarize first several of McKinnon's key propositions and prescriptions and then review these in a critical spirit.

Proposition 1: the belief that real exchange rates affect the current account is a "false doctrine." Real expenditure changes triggered by exchange rate movements offset gains in competitiveness. Current account imbalances should be adjusted by fiscal policies, not real exchange rates.

Proposition 2: fixed exchange rates are superior to flexible exchange rates as a means of adjusting to real disturbances and to achieve price stability.

■ *Rudiger Dornbusch is Ford International Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts.*

Proposition 3: purchasing power parity (PPP) is a good guide to equilibrium exchange rates.

Proposition 4: exchange rate movements and volatility are brought about primarily by money demand shifts between bank accounts at home and abroad.

Proposition 5: fixed nominal exchange rates are socially preferable to flexible rates. In fact, fixing exchange rates compensates for the lack of complete markets which McKinnon calls "Arrow-Debreu conundrum."

Proposition 6: world monetary growth should be targeted to achieve price level stability. Countries should follow an assigned domestic credit target and use symmetric, unsterilized intervention to stabilize exchange rates.

No Effects From Real Exchange Rates

The most startling assertion is that real exchange rates have no effect on the current account. The view is extraordinary because it runs so much counter to what one thinks the U.S. experienced in the 1980s. The argument is readily countered by looking at the determinants of net exports. The equation below reports a regression of net exports (as a fraction of GNP) on the real exchange rate, relative expenditure levels and a time trend, using quarterly data for the United States for the period 1975:1–1986:4. (For a more detailed discussion, see Dornbusch, forthcoming.)[1]

$$\begin{array}{llllll} \text{NET} = & 90.0 & -5.96\log(P/eP^*) & -12.23\log(D/D^*) & -0.04\,\text{Time} \\ & (7.86) & (-3.78) & (-3.68) & (-6.18) \end{array}$$

$$R^2 = 0.96 \quad \text{Rho} = 0.43 \quad \text{SER} = 0.28$$

where NET is nominal U.S. net exports (NIA) as a fraction of GNP,

P/eP^* is relative value added deflator in manufacturing,

D/D^* is relative levels of real total spending.

The regression shows that real exchange rates and relative spending levels are significant determinants of net exports. A 16 percent real depreciation shifts net exports by 1 percent of GNP as does an 8 percent rise in the relative level of foreign real spending. There is an adverse time trend in net exports which leads over a six-year period to a deterioration in net exports by 1 percent.

But McKinnon's argument is not vulnerable to the evidence offered here. He argues that real exchange rate changes are ineffective because they trigger compensat-

---

[1] The real exchange rate measure is the IMF's relative value added deflator in manufacturing reported in *International Financial Statistics*. The index of the relative levels of real gross domestic spending was calculated as follows: for the U.S. real gross domestic spending was used and for the rest of the world a weighted average index for the non-U.S. OECD constructed from interpolated annual data. The weights are the OECD GNP shares.

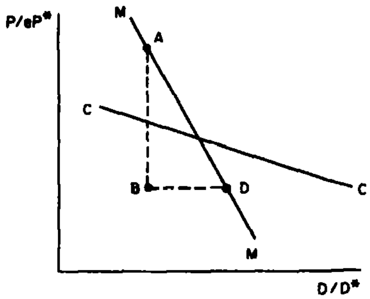


Fig. 1. Relative prices, spending and external balance

ing movements in demand, so that the combined effect is to leave current account balances unchanged. Figure 1 makes the point. On the vertical axis we show the real exchange rate denoted by P/eP^* and on the vertical axis relative levels of spending, D/D^*. The schedule *CC* shows a zero current account using, for example, the above equation. McKinnon argues that there is another relation, *MM*, which links relative levels of demand to the real exchange rate. Now consider a real depreciation from point *A* to *B*. Clearly the real depreciation will lead to a surplus, but it also triggers a change in relative spending levels leading to point *D*. Still, a surplus remains at *D*, although smaller than at *B*. Thus, everything hangs on the empirical shape of the *MM* schedule.

What is the evidence of the past five years? During the period of dollar appreciation, U.S. real spending increased much more rapidly than spending abroad. Since 1985–86 the reverse is true; real expenditure has grown more rapidly abroad than in the United States. This would suggest that the *MM* schedule would be positively sloped (or shifting to the right), reinforcing rather than offsetting the real exchange rate effects. In fact the *MM* schedule depends significantly on policies. Of course, a government can use demand policies to frustrate the impact of real exchange rate changes. But a general dominance of income effects is most definitely counter to all evidence from trade equations.

In sum, real exchange rates do matter for the current account and real exchange rate movements can help adjust external balances by bringing about old-fashioned expenditure switching. I agree with McKinnon on the need for fiscal correction in the United States and I also agree that the fiscal correction when it ultimately comes will help correct the trade deficit. But there is a twin objective to be met: full employment and external balance. Fiscal policy cannot do everything; that is why God invented the real exchange rate.

Adjustment to Real Shocks

McKinnon tells us that the United Kingdom experienced appreciation and a loss of manufacturing competitiveness and sharp disinflation because of a favorable oil

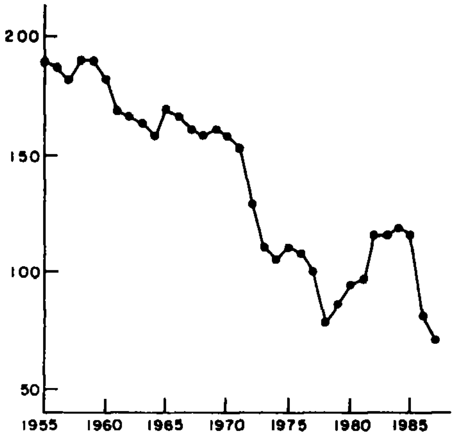


Fig. 2. U.S.-Japan relative price levels (index 1980 = 100)

shock in 1979-81. Had rates been fixed these effects could have been avoided or at least the adjustment would have been more gradual.

Surely the U.K. authorities, had they wished to do so, could have avoided appreciation by expanding money. They did not do so, presumably, because they were anxious to achieve a disinflation via currency appreciation. Whatever the wisdom of that decision, with an inflation rate of 18 percent in 1980 there is little surprise that sharp disinflation was viewed as a policy triumph, not a shortcoming. No country need experience real appreciation, except when it wishes to reduce inflation and for that reason takes advantage of a speculative inflow to bring about imported disinflation. Managed flexible rates thus seem superior to fixed rates because they allow a choice in the trade-off between growth and inflation. A sufficiently sizeable monetary expansion will always kill an appreciation and the monetary-fiscal mix can deal with the optimal, non-inflationary absorption of favorable or unfavorable shocks.

Now suppose an unfavorable shock, such as a loss of export markets. To sustain full employment there needs to be a real depreciation. That conclusion is routinely taught in trade theory and presumably carries over to open economy macroeconomics. It is easier to achieve real depreciation by exchange rate depreciation or by cutting money wages? Exchange rate movements seem the easier way.

Even if one does not want to accept boundless exchange rate flexibility, there is here at least a strong argument for trend flexibility in real exchange rates. Figure 2 shows the U.S.-Japanese real exchange rate (measured by relative GNP deflators in a common currency) for the past thirty years. Clearly the real exchange rate has gradually moved over time. Should this trend of real appreciation be accomplished by divergent money wage trends? If so, national inflation rates will differ even under fixed rates, as we will now argue.

PPP and Common Inflation

With fixed exchange rates but differing rates of productivity growth there may be very strong divergences in inflation rates. If exchange rates were fixed, who should get zero inflation? The slowly or the rapidly growing country?

Exchange rate movements may be required for one of two reasons. The first is to offset divergent trends in national price levels. For example, imagine two economies, each with income equal to spending, without structural change and with differing rates of inflation. A flexible exchange rate would be depreciating at a rate equal to the inflation differential. That is the obvious role for exchange rate flexibility.

When one country experiences productivity growth at a rate higher than the other they cannot have the same rate of inflation. (See Kravis and Lipsey, 1983, for a discussion of the implications of differential productivity growth for relative price levels.) Suppose that the productivity growth is concentrated in tradeables (manufactures, for concreteness). Under fixed exchange rates and a policy of zero inflation for tradeables, both countries will experience inflation, but it will be higher in the country with the higher rate of productivity growth. Consider a concrete example: there are two countries, the United States and Japan, each with 1 percent productivity growth in nontraded goods, Japan with 6 percent productivity growth in tradeables and the U.S. with only 2 percent. For zero inflation of tradeables, wages in Japan must rise at 6 percent per year and in the United States at 2 percent. Let the nontraded goods sector account for 50 percent of the consumer price index in each country. U.S. inflation (equal to nontraded goods cost increases weighted by the 0.5 share) will be 0.5 and in Japan it will be 2.5 percent. Thus if productivity growth differentials are significant a problem emerges for the joint planning of world inflation. It might be argued that the inflation differential stemming from productivity growth is very small and hence presents an insignificant obstacle. But looking back to the 1960s, even these small differentials were unacceptable.

McKinnon tells us that at 180 yen to the dollar, given the current level of prices, a dollar buys broadly the same basket of goods in Japan and in the United States; hence 180 yen per dollar is the equilibrium exchange rate. What an extraordinary resurgence of a doctrine long thought dead and buried by Viner, Samuelson and many other writers on the issue!

Here is the simple argument against PPP: U.S. and Japanese goods are not perfect substitutes and hence issues of the law of one price or spatial arbitrage are irrelevant. Imagine that at the PPP exchange rate U.S. firms do not find buyers for their basket while Japanese firms do overtime. Something will have to give, like the real exchange rate.

When a country opens new markets, introduces new products or experiences a favorable price shock for its traditional exports, the real exchange rate will adjust. The thought that PPP governs exchange rates is only appropriate to a world where equilibrium relative prices never change. If they change a lot because real disturbances are large, then PPP is best forgotten. The extraordinary manufacturing advances in Japan certainly are the kind of disturbances that would have led Cassel to concede the need for an equilibrium departure from PPP.

Determinants of Exchange Rates

In McKinnon's view, exchange rate movements are due predominantly to unanticipated monetary disturbances and, more narrowly, to unanticipated shifts in money demanded from one country's M_1 to another's. This currency substitution occurs in response to volatile shifts in inflationary expectations fed by political and financial real events. Because central banks do not use exchange rate targeting as their money supply principle, exchange rates will be tossed around by these unanticipated money demand developments.

In earlier work McKinnon has advocated this money demand view of exchange rate movements more strongly. (See the discussion in Cuddington (1983), Dornbusch (1986, 1987a), Spinelli (1985) and McKinnon (1982, 1984a, b).) Now he recognizes that portfolio disturbances involve dominantly shifts between interest bearing assets. But the basic emphasis on an M_1-disturbance view of exchange rates is still lingering even in the present version of his proposal when he argues: "Because private portfolio preferences are now so volatile across dollar, yen and mark assets any one central bank cannot now easily estimate the demand for its own money."

This narrow view of exchange rate determinants leaves out two important explanations for exchange rate movements. First, central banks have objectives other than the exchange rate. They may know perfectly well the response of money demand to, say, a tax cut, but do not choose to accommodate because they are concerned to stabilize activity or inflation. Second, international portfolio shifts are taking place between one country's bonds and another's. Central banks may recognize this but fail to intervene with appropriate sterilized intervention because they welcome for stabilization purposes the movements in real exchange rates or because they feel they cannot identify the source of disturbance. Not knowing whether they face a real or a portfolio disturbance, and in the latter case what kind, the authorities allow exchange rate movements to occur without intervention. Accordingly exchange rates can move for any of a number of reasons. I am not aware of evidence that would support the view that money demand disturbances are the dominant source.

Exchange Rate and the Current Account

If exchange rate movements are not caused primarily by shifts in money demand from one country's M_1 to another's, there is no basis for preferring fixed exchange rates over flexible rates. Moreover, there certainly is no presumption for using unsterilized intervention as the rule.

The analysis by Poole (1970) of the choice between interest rate pegging and a monetary aggregate target showed that optimal monetary policy should offset or accommodate purely financial disturbances, thus avoiding unnecessary spillover to real activity. But Poole noted that monetary policy risks misreading changes in interest rates as money demand disturbances when they in fact are the result of real disturbances. That risk is why outright interest rate pegging is almost never appropriate. The same argument applies to exchange rate pegging. If there are real disturbances, then pegging the exchange rate by intervention is inappropriate. When disturbances originate outside the money market, unsterilized exchange rate pegging risks magnifying the initial disturbance in other markets.

An important implicit assumption of the case for fixed exchange rates is that countries can and will agree on coordinating fiscal policies. If they do, and if the coordination leads to no major and asymmetrical swings in full employment budgets, then of course one important source of exchange rate movements has disappeared. With fiscal disturbances to the exchange rate eliminated, a monetarist treatment is already more (though not fully) justified. But there is, of course, no reason to believe that fiscal policies will be coordinated. On the contrary, all evidence indicates that legislatures and executive branches of government are completely unwilling to let even domestic policy mix considerations impinge on budget decisions. If fiscal coordination cannot be taken for granted, one cannot simply go ahead and announce the desirability of fixed exchange rates with monetary rules.

If uncoordinated, large fiscal swings are a possibility, there is no presumption that fixed rates with unsterilized intervention are the best kind of monetary policy. Our fiscal expansion would now lead to an expansion of home credit and to a contraction abroad. World interest rates would rise and our trading partners risk falling into recession. Neither the fixed rate nor the monetary rule seem in this context a very attractive feature. Of course, that brings out precisely the problems of the past five years.

Concluding Remarks

I would like to emphasize two further points. First, in discussions of the exchange rate regime there has been too little emphasis on differences in economic philosophy. Europe and Japan are significantly more monetarist than the United States; when the chips are down, they prefer recession or depression to inflation. If our preferences do not run that way the union is flawed from the outset. This is quite apparent in the present context of interest rates. Short term real interest rates continue to be extremely high. Europe and Japan are obviously suffering from the dollar decline, but even now they are reluctant to reduce their real rates to stem further dollar depreciation, fearing that even near-zero inflation is a big issue.

My second point concerns trend changes in equilibrium real exchange rates between Europe, the U.S. and Japan that result from the emergence of the newly industrialized countries. As Korea, Brazil and other trading countries become a major force in world trade there will need to be pervasive adjustments in relative prices. The next ten years will bring much of that. What a poor idea to start that process by

removing exchange rate flexibility and shifting the burden onto inflexible wages. Debreu's conundrum may apply to goods markets under flexible rates, but it is certainly worse for labor markets.

References

Cuddington, John, "Currency Substitution, Capital Mobility and Money Demand," *Journal of International Money and Finance*, August 1983, *2*, 111-133.

Dornbusch, Rudiger, *Dollars, Debts and Deficits*. Cambridge, MA: MIT Press, 1986.

Dornbusch, Rudiger, "Exchange Rate Economics: 1986," *Economic Journal*, March 1987a, *95*, 1-18.

Dornbusch, Rudiger, "External Balance Correction: Depreciation or Protection", *Brookings Papers on Economic Activity*, forthcoming.

Kravis, Irving and Robert Lipsey, "Toward an Explanation of National Price Levels," International Finance Section, Princeton University, Princeton Studies in International Finance, No. 52, 1983.

McKinnon, Ronald I., "A New Tripartite Monetary System or a Limping Dollar Standard," International Finance Section, Princeton University, Princeton Essays in International Finance, No. 106, 1974.

McKinnon, Ronald I., "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, *70*, 320-33.

McKinnon, Ronald I., "An International Standard for Monetary Stabilization, Institute for International Economics, Policy Analyses in International Economics, No. 8, 1984a.

McKinnon, Ronald I., "How to Coordinate Central Bank Policies," *New York Times*, July 12, 1984b.

McKinnon, Ronald I., "Currency Protectionism: Parity Lost," *Wall Street Journal*, February 2, 1987a.

McKinnon, Ronald I., "A Gold Standard Without Gold. Putting a Stop to Rapid Currency Swings", *New York Times*, April 12, 1987b.

Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, *LXXXIV*, 197-216.

Spinelli, Franco, "Currency Substitution, Flexible Exchange Rates, and the Case for International Monetary Cooperation." International Monetary Fund, Staff Papers, December 1985, *32*, 755-783.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 113–119

Comment on McKinnon's Monetary Rule

John Williamson

If the world wishes to return to fixed exchange rates, then McKinnon's monetary rule provides the right way to go about the task. Over the past year he has modified his longstanding proposals in order to accommodate the breakdown of monetarism: he no longer urges the participants to follow a *fixed* rate of aggregate monetary growth consistant with zero inflation, but instead to *vary* their aggregate monetary growth in a way that will achieve zero inflation, thus recognizing the variability in velocity. Similarly, he has widened his recommended band to 10 percent (at least initially) and now calls for interest rate variations to be used to manage rates except when they move to the margins, at which point symmetrical unsterilized intervention becomes mandatory. I see no reason to quibble with his analysis on the means to fix exchange rates.

The remaining question is of course whether the *end* of fixed nominal exchange rates is a sensible one. In questioning that, I am not challenging his analysis of the costs that floating rates have imposed on the world[1] or his judgment of their vast magnitude. I have tried to argue the same myself (Williamson, 1985, pp. 38–46). But that does not mean that there are not real social benefits from exchange rate flexibility. In this note I shall outline those benefits, and then proceed to describe my principal analytical disagreement with McKinnon.

The alternative which I shall contrast to the McKinnon Rule is the extended target zone proposal, developed in Edison, Miller and Williamson (1987) and sum-

---

[1] The only qualification I would offer is that options markets of longer maturity than currently function might enable producers to lay off much of the risk involved in the investment decision that McKinnon suggests to be inherently uncoverable.

■ *John Williamson is Senior Fellow, Institute for International Economics, Washington, D.C.*

marized in Williamson (1987). This proposal involves a comprehensive set of rules for policy coordination among the main industrial countries. Each participating country would have an endogenous target rate of growth of nominal income $\hat{y}^*$, set equal to the estimated rate of growth of productive potential g, plus some fraction of the inherited rate of inflation, $\hat{p}_{t-1}$ (to implement a gradualist disinflation strategy), plus a positive function of the deflationary gap d:

$$\hat{y}^* = g + \alpha\hat{p}_{t-1} + \beta d.$$

This would provide one intermediate target for each country. The other would be a target for the (real effective) exchange rate, where the target would be set at a level estimated to reconcile internal and external balance in the medium term.

This set of $(2n - 1)$ intermediate targets would be pursued by the following set of assignment rules: (1) the *average* level of world interest rates would be revised up (down) if aggregate growth of nominal income is threatening to exceed (fall short of) the aggregate target growth of nominal income for the participating countries; (2) *differences* in interest rates among countries would be revised when necessary to prevent exchange rates deviating from their target levels by more than, say, 10 percent; (3) national *fiscal policies* would be revised with a view to achieving national target rates of growth of nominal income.

Rule (3) endorses Keynesian fiscal policy. Rule (2) embodies the target zone proposal. (Incidentally, these interest rate adjustments might well be supplemented by exchange market intervention, which could alternatively be described as the use of unsterilized intervention, or even replaced by sterilized intervention when market pressures are limited.) Rule (1) aims at assuring (in a manner inspired by McKinnon) that aggregate monetary policy should be guided by the collective need of the participating countries. In fact, Rule (1) may be termed "McKinnon without the monetarism."

The set of "assignment rules" suggested above should be interpreted as policy guidelines rather than as inflexible commitments, for two reasons. The first is that it would be an enormous advance to persuade sovereign states that their policies ought normally to be designed not just to further their own interests but also in a manner consistent with the needs of the rest of the world economy, and to seek to deny countries the ultimate right to choose their own policies would certainly jeopardize the political acceptability of the proposal. The second is that the rules would need to be interpreted with a degree of flexibility. For example, Rule (1) might perpetuate an undesirably lopsided mix of tight money and lax fiscal policy if it were not supplemented by a joint resolve to superimpose a secular tightening of fiscal policy on Rule (3) when real interest rates seemed to be stuck above their historical norm. Indeed, an individual country with an undesirably large budget deficit should combine Rule (3) with a trend toward fiscal tightening. To take another example, Rule (1) should certainly not be interpreted to require central banks to push interest rates down aggressively against market resistance.

The Social Functions of Exchange Rate Flexibility

A first important function of exchange rate flexibility is that of reconciling differential inflation. McKinnon will deny that there is any need to have differential inflation, except in countries with a fiscal requirement for the inflation tax, since we now know that the long-run Phillips curve is vertical and hence that faster inflation does not buy worthwhile output gains. This seems to me too simplistic. It is one thing to agree that the appropriate inflation target is zero and that it should be taken seriously. It is another to feel confidence that such a target will always be achieved.

Suppose that a participating country experiences a burst of inflationary pressure from any cause other than a sudden urge of the central bank to print money (which would of course be repressed by the McKinnon Rule). Arbitrage is not sufficiently perfect to prevent prices rising in the afflicted country, as shown by McKinnon's (1979) work involving the distinction between differentiated manufactured products, as opposed to homogeneous primary commodities. The inflation comes to an end when the joint effect of (imperfect) arbitrage pressures in tradables and the recession induced by the loss of competitivity in tradables is sufficient to offset the initial inflationary thrust. This process necessarily involves exposing the tradables sector to the sort of pressures that McKinnon so rightly deplores in the second paragraph of his paper. Hence only those who are convinced that inflation can originate in no way except by printing money[2] should accept the McKinnon route to restoring price stability.

The nominal income target in the extended target zone proposal provides an alternative (Wicksellian) strategy for restoring price stability. Even if the real exchange rate were maintained approximately constant by adjustments in the nominal rate designed to neutralize differential inflation, the deflation required to prevent the nominal income target being exceeded would bring inflation to an end, while generalizing the necessary deflationary pressures over the whole economy rather than concentrating them on tradables. (If some modest appreciation is needed to equalize the pressure on tradables and nontradables, that is allowed by the wide band.) This strategy seems to me vastly preferable to McKinnon's, since it avoids the pressures that produced deindustrialization in the United States over the years of the high dollar (as in Britain in the early 1980s or Argentina in 1979–81).

A second social function of exchange rate flexibility is that of facilitating payments adjustment when this proves necessary—for example, in response to "permanent" real shocks—by changing the incentives to export and import. My main analytical disagreement with McKinnon, on which I dwell later, concerns his implicit theory of payments adjustment, which leads him to reject any role for exchange rate changes in facilitating that process. Suffice it to note here that in his discussion of "The Transfer Problem" he does recognize that even under fixed exchange rates a flow of capital from Japan to the United States would involve "some (slow) bidding up of nontradable prices (largely services) in the United States... and a slower rate of increase in Japanese nontradables prices... Although," he adds, "these relative price movements within both countries would be modest... as with past experiences with fixed exchange rate regimes approximating a common currency area." Some of us think that it was because these movements were so modest as to result in no adjustment that Bretton Woods broke down.

---

[2] Presumably everyone agrees that perpetuation of inflation requires monetary accommodation, but that is not the point at issue.

A third function of exchange rate flexibility is to liberate monetary policy to pursue interest rate targets at variance with those in the rest of the world. If one country is suffering a deeper recession than its partners, it may legitimately wish to ease monetary policy relative to other countries, and that will be feasible only if its currency can depreciate so as to create an expectation of a subsequent rebound that will compensate investors for the temporarily low interest rates. Conversely, a country with abnormally severe inflation may legitimately seek to raise interest rates temporarily, which will require an appreciation. A wide band within which exchange rates are allowed to move around parity provides scope for such temporary variations in monetary policy to pursue anticyclical objectives. (As it happens, the definitive analysis of this issue was provided by McKinnon, 1971.)

The final legitimate function of exchange rate flexibility is that of absorbing a part of speculative pressures. Instead of requiring that every change in speculative sentiment lead to a change in international reserves and/or interest rates, one can allow changes in the exchange rate to take some of the strain. Provided these changes do not lead to the prolonged and substantial movements away from equilibrium that constitute misalignments, they do little harm.

In his discussion of my proposal for target zones, McKinnon charged that the option of occasional changes in the target real effective exchange rate, designed to facilitate adjustment, would expose a target zone system "to speculative attack as international investors tried to guess how exchange rates in the future might differ from those now prevailing," as under the Bretton Woods system. In fact, the target zone proposal envisages bands more than wide enough to preempt any danger of a return to the speculative crises of Bretton Woods. There is a fundamental difference between an adjustable peg and a crawling peg in terms of their consistency with asset market equilibrium, and therefore their vulnerability to speculative pressures. The analytical distinction between an adjustable peg and a crawling peg is whether the band width is less or greater than the size of parity changes. A target zone system falls in the category of a crawling peg, since it is highly unlikely that markets would come to expect changes of target exchange rates of more than 10 percent (let alone 20 percent). Hence expectations could be continuously satisfied.

A target zone system would permit exchange rate flexibility to fulfill all four of its genuine social functions. It could reconcile differential inflation, by virtue of the provision for automatic adjustment of the nominal target to maintain the target zone constant in real terms. It could facilitate payments adjustment, by changing the real zone in response to permanent real shocks. It would permit a degree of independence for anticyclical monetary policy, by virtue of the wide band. And it could absorb speculative shocks, through the wide band. What target zones do *not* do is rely upon the fantasy that markets are better than governments at thinking about where exchange rates ought to be—a fantasy because the time horizon of most market operators is about 5 hours, whereas it needs to be about 5 years if exchange rates are to be set at levels consistent with satisfactory macroeconomic performance.

Despite this substantial element of flexibility, target zones could alleviate both of the great inadequacies from which floating has been revealed to suffer—misalignments, and the lack of pressure to coordinate policies among governments. While a zone of ± 10 percent is wide, it would have constrained much of the erratic variation that has been experienced under floating, and as shown in McKinnon's Figure 1. And of course, the incorporation of the proposal in the comprehensive scheme of policy coordination already sketched illustrates how far this proposal is from the laissez-faire policy of 1973–85. Even if it were to be adopted initially without nominal income targets and fiscal commitments, it could be expected to exert pressures toward coordination in those areas.

The Analysis of Payments Adjustment

Conventional theory argues that both exports and imports, and therefore the current account, are functions of income/absorption levels at home and abroad and of the relative price of domestic inputs/output in terms of foreign inputs/output, with lower domestic prices ultimately leading to bigger net exports (provided the Marshall–Lerner condition is satisfied). A domestically generated increase in income/absorption is expected to lead to a reduction in net exports of some fraction of the initial change, because the marginal propensity to import is a fraction. These are all partial equilibrium relationships which have to be embedded in some appropriate general equilibrium model before one can draw conclusions about the ultimate impact on the current account of such changes as a reduction in the U.S. budget deficit or a depreciation of the dollar.

McKinnon claims that the U.S. "trade deficit of $150 billion to $200 billion a year merely reflects the savings-investment gap in the American economy created by the not coincidentally equally large U.S. fiscal deficit" and that "getting rid of this chaos in the American public finances is the only satisfactory way by which the trade deficit can be reduced" since "exchange rate changes have no predictable effect on net trade balances." These statements appear to imply that he believes that changes in the budget deficit are always reflected one-for-one by changes in the trade balance whatever the level or change in relative prices. The condition for this to be true with constant prices is that the marginal propensity to import multiplied by the Keynesian multiplier be equal to unity.[3] Since the marginal propensity to import is probably around 0.2 and the multiplier around 1.5, this hardly appears likely.

The conventional analysis argues that what is needed to correct the trade balance is a combination of more competitive relative prices, which can be induced by depreciation provided that inflation is prevented from completely neutralizing the initial change, and a cut in absorption. It is indeed possible that a depreciation

---

[3] A conventional import function includes the term $\Delta M = m\,\Delta Y$ (plus the effect of changes in relative prices), while $\Delta y = k\,\Delta D$ where k is the Keynesian multiplier, m is the marginal propensity to import, ΔD is the change in the budget deficit, ΔM is the change in imports and ΔY is the change in income.

unaccompanied by other measures might increase the trade deficit, if an undervalued dollar touched off an investment boom in the United States that increased imports and an overvalued yen depressed investment and hence led to recession in Japan. But this is not an argument for correcting the whole deficit by income deflation, since this would certainly require a recession in the United States. It is an argument for combining the dollar depreciation with a reduction in the U.S. fiscal deficit and, given the weakness in the world economy, a fiscal stimulus in Germany and Japan and other surplus countries. I find it depressing that it remains necessary to spell out such an elementary point at such length 36 years after Meade's (1951) definitive analysis (which was a general equilibrium analysis that it is a caricature to describe as the "elasticities approach to the balance of trade").

McKinnon's view that current account balances change one-for-one with changes in the fiscal stance is a reincarnation of the doctrine propagated by the New Cambridge school in Britain in the early 1970s. This view would be valid if, but only if, the gap between private savings and investment were constant, which it most certainly is not (Marris, 1985, Figure 1.1). Moreover, one wonders whether this proposition is supposed to be affected by changes in the foreign fiscal stance. If the rest of the world applies an equal fiscal stimulus as the United States restores fiscal discipline, is the U.S. current account balance supposed to improve twice over? Or is the action of the rest of the world irrelevant?

Yet it is McKinnon's eccentric view that the U.S. current account balance changes one-for-one with changes in the U.S. fiscal stance that is responsible for his espousal of the PPP criterion as the basis on which to estimate "correct" exchange rates. With a more conventional view, one recognizes that the desirable level of competitiveness depends, for example, on the size of transfer that it is desired to effect, and on real factors such as the price of oil. There is no basket of goods whose price necessarily has to be equal in the two countries in order to achieve macroeconomic equilibrium. Nor, for that matter, will equal inflation on any measure generally preserve macroeconomic balance at a fixed exchange rate in the presence of real shocks, including the need to make a different real transfer.

In fact, the McKinnon and Ohno (1986) criterion for identifying a "correct" PPP, on which McKinnon relies, does not involve either equality of the prices of a particular basket of goods nor some base period being identified as a period of equilibrium. Essentially what they do is search for that exchange rate which exerts neither inflationary nor deflationary pressure on the price level; that is, at which actual inflation is equal to domestically generated inflation. They interpret imported inflation as evidence of an undervalued exchange rate, and vice versa. The problem with this criterion is that ultimately an economy may adapt to a wide range of values of the real exchange rate. Income earners may come to acquiesce in depressed real incomes and inflation may disappear despite an undervaluation (in the sense defined in the next paragraph). Subsequent appreciation may then lead to falling prices for a time, even though after a time income aspirations will adjust and a quite different real exchange rate would then be identified by the McKinnon-Ohno PPP criterion. In other words, if imported inflation depends on *changes* in the exchange rate rather than its deviation from some PPP level, their criterion is hopelessly ambiguous.

However, the alternative to the PPP criterion for identifying "correct" exchange rates should most certainly not be labelled the "balanced trade" principle, for two reasons. In the first place, its application (as in Williamson, 1985) does not require pursuit of balanced trade: some agreed set of target current account balances is required as an input, but these need not be zero balances, nor imply zero trade balances. Second, and more fundamental, the alternative principle involves seeking *macroeconomic balance*—an exchange rate consistent with simultaneous achievement of internal and external balance in the medium run. It should accordingly be labelled the "macroeconomic balance" or "internal-external balance" criterion, rather than the "balanced trade criterion."

Conclusions

While McKinnon's criticisms of the regime of floating exchange rates are almost entirely justified, I have argued that both his analysis and his diagnosis are flawed. His analysis presupposes a very special set of parameters which result in changes in the budget deficit leading to equal changes in the trade balance while leaving output unchanged. In fact, achieving this result requires also changes in relative prices, which are best induced by changes in exchange rates. This points to one of the four important social functions of exchange rate flexibility that are denied or ignored by McKinnon and that would be forfeited under his proposal. In contrast, target zones offer escape from the major costs of floating while retaining the important advantages that flexible rates offer.

■ *The author acknowledges helpful comments on a previous draft by Shafiqul Islam and Stephen N. Marris.*

References

Edison, Hali, Marcus Miller, and John Williamson, "On Evaluating and Extending the Target Zone Proposal," *Journal of Policy Modeling*, Spring 1987, *9*, 199–224.

Marris, Stephen N., *Deficits and the Dollar*. Washington: Institute for International Economics, 1985.

McKinnon, Ronald I., *Monetary Theory and Controlled Flexibility in the Foreign Exchanges*, Princeton Essays in International Finance No. 84, 1971.

McKinnon, Ronald I., *Money in International Finance: The Convertible Currency System*. New York: Oxford University Press, 1979.

McKinnon, Ronald I., and Kenichi Ohno, "Getting the Exchange Rate Right: Insular Versus Open Economies," unpublished paper, Economics Department, Stanford University, 1986.

Meade, James E., *The Balance of Payments*. London: Oxford University Press, 1951.

Williamson, John, *The Exchange Rate System*. Rev. ed. Washington: Institute for International Economics, 1985.

Williamson, John, "Exchange Rate Management: The Role of Target Zones," *American Economic Review*, May 1987, *77*, 200–204.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 121–144

Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics

Dale J. Poirier

> *"It is difficult for me to tone down the missionary zeal acquired in youth, but perhaps the good battle is justified since there are still many heathens."*
>
> I. J. Good (1976, p. 126)

The "heathens" to whom I. J. Good refers are the non-Bayesians of the world and they are my targeted audience as well. Specifically, I plan to discuss, in as simple and nontechnical a fashion as possible, the subjectivist-Bayesian attitude toward model building in econometrics and to contrast it with the standard frequentist attitude. Most of the issues I raise are familiar to statisticians but not to economists. Rather than give the suspicious reader a menu of Bayesian techniques, I hope to create an interest in acquiring a taste for the Bayesian cuisine by recommending five pragmatic principles.

While neither the Bayesian nor the non-Bayesian flocks are homogeneous with respect to the beliefs of their own members, I believe the "within-group" differences

■ *Dale J. Poirier is Professor of Economics, University of Toronto, Toronto, Ontario, Canada.*

are small relative to the "between-group" differences, and hence I focus on the latter.[1] To convey what I believe is the principle distinguishing attitude between these two groups, I refer to their respective positions as "subjectivist" and "frequentist." The basic differences between these positions arise from different interpretations of "probability."

Frequentists interpret probability as a property of the external world, i.e., the limiting relative frequency of the occurrence of an event as the number of suitably defined trials goes to infinity. According to this view a "true" model or data generating process exists involving parameters which attain an objective existence.[2] Familiar introductory textbook examples of such parameters are the "probability of heads" in coin flipping and the "length" of some physical object (e.g., a farmer's field) subjected to repeated measurements.

I am not sure, however, that such simple examples are representative of the models of human behavior that economists use, nor am I sure that the parameters which appear in economists' models are endowed with an objective existence like the length of a field. When pressed, I think few econometricians would argue that "utility functions" or "representative agents" exist in the world. Rather I think they would feel more comfortable with the view that such entities are theoretical constructs that give rise to models which provide useful windows through which to view the world, and by which many observables in the world (like prices and quantities) can be related to each other.

For a subjectivist, probability is interpreted as a degree of belief fundamentally internal to the individual as opposed to some characteristic of the external world. Subjective probability measures a relationship between the observer and events (not necessarily "repetitive") of the outside world, expressing the observer's personal uncertainty about those events. The subjectivist paradigm is designed to produce "coherent" revisions in beliefs about future observables in light of observed data.

This personal uncertainty can be characterized in terms of an individual's (e.g., an economic forecaster's) willingness to engage in buying and selling lottery tickets involving events that are potentially observable (e.g., the yen/dollar exchange rate thirty days hence).[3] In such a setting the probability of an event is defined to be the certain gain a risk-neutral individual judges to be equivalent to a unit gain obtained if the event occurs. The only restriction to be placed on such personal degrees of belief is that the individual never assigns probabilities to events in such a way that an opponent can choose wagers which guarantee that the individual loses no matter what the outcome of the lottery (known as making "Dutch book"). This restriction, called *coherency*, guarantees that such value assignments satisfy the usual axioms of probability, except that the additivity axiom holds only for finite (rather than countable) unions of events.

---

[1] For example, based on eleven facets, Good (1971) argues there are at least 46,656 types of Bayesians. Here I treat all Bayesians in their "purest" form, namely, as subjectivist.

[2] In this regard see Dawid (1982a, pp. 217–218).

[3] This lottery-based formulation is due to de Finetti (1937). For a critical discussion of it from an economist's perspective, see Leamer (1986).

While this betting scenario provides a potentially operational formulation of the subjectivist position, I wish instead to emphasize here its role as an ideal in which the researcher is imagined in a decision making setting not unlike that commonly envisioned for the agents in an economic theory. The decisions facing the researchers regard forecasts, estimates and inferences. Coherency seems an a priori reasonable requirement to impose on beliefs in order to provide a sense of "rationality" on the resulting decisions. Simply put, researchers should demand of themselves what they ask of the actors in their models.

Before closing this brief introduction to the frequentist/subjectivist dichotomy of views on probability, let me add that there exists a "subjective" element in any interpretation of probability. To calculate a relative frequency it is necessary subjectively to define the class of events (known as a *collective*) over which to count the frequency. The fact that different individuals may agree on the choice of a collective does not make it any less subjective. Frequentists like to think of "probability" as a property of the external world, but the choice of a collective brings the individual into the determination.

The Likelihood Principle

Basic Development

While I perceive the choice between the frequentist and subjectivist perspectives decidely to favor the latter directly on fundamental grounds, it is possible to approach the choice indirectly by assessing each perspective's adherence to other basic principles. This latter approach, advocated by Berger (1986) among others, is followed in this section. For pedagogical reasons, but without loss of generality, I focus the discussion on a familiar problem (Example 1) in order to introduce important concepts and notation.[4] So as to not prejudice the case, the presentation is in an objectivist tradition.

Example 1A. Let $Z_1, Z_2, \ldots$ be a sequence of independent and identically distributed Bernoulli random variables taking on the values "1" and "0" with probabilities θ and $1 - \theta$, respectively, where θ is an unknown parameter satisfying $0 \leq \theta \leq 1$. For concreteness the reader may wish to think of this sequence as corresponding to the outcomes of independent identical flips $t = 1, 2, \ldots$ of a coin resulting in trial t in either a "success" (say a "head"), implying $Z_t = 1$, or a "failure" (say a "tail"), implying $Z_t = 0$. Consider two possible *experiments* (i.e., data generations), denoted E_1 and E_2, giving rise to such a sequence. According to E_1 a sample of size T_1 is generated directly, where sample size T_1 is predetermined. According to E_2 sampling continues until m "failures" are obtained, where $m > 0$ is a predetermined integer. In the case of the second experiment the sample size T_2 is determined by the sampling outcomes, and the T_2 observations consist of $T_2 - m$ "successes" and m "failures". Let Y_1 and Y_2 denote, respectively, the number of successes obtained in each experiment. Y_1 and Y_2 are random variables and their realized values in a particular running of the experiment are denoted by y_1 and y_2, respectively. From introductory statistics it is known that the distributions of Y_1 and Y_2 are binomial and negative binomial, respectively, with probability mass functions

---

[4] For a complete discussion, see Berger and Wolpert (1984a) and Berger (1985b).

$$f_1(y_1|\theta) \equiv \text{Prob}(Y_1 = y_1|\theta) = \frac{T_1!}{y_1!(T-y_1)!}\theta^{y_1}(1-\theta)^{T_1-y_1}, \qquad y_1 = 0,1,\ldots,T_1$$

$$f_2(y_2|\theta) \equiv \text{Prob}(Y_2 = y_2|\theta) = \frac{(y_2+m-1)!}{(m-1)!y_2!}\theta^{y_2}(1-\theta)^{m}, \qquad y_2 = m, m+1,\ldots$$

Experiment $E_J(J = 1,2)$ is fully characterized by the triple $\{Y_J, \theta, f_J(y_J|\theta)\}$. It is important to note that θ is the same for both experiments. The *likelihood function* $L_J(\theta; Y_J)(J = 1,2)$ corresponding to E_J is defined to be the probability mass function $f_J(y_J|\theta)$ viewed as a function of θ for the realized value y_J. Given the sample outcome y_J, $L_J(\theta; y_J)$ conveys the notion of how "likely" is a particular value θ. Although the exact interpretation of "likely" is unclear, it definitely should *not* be interpreted as "probability." Finally, the *evidence* about θ arising from experiment E_J and realized data y_J, denoted $Ev[E_J, y_J]$, is defined broadly to be any inference, conclusion or report concerning θ based on E_J and y_J. Common examples of "evidence" are point estimates, interval estimates or the results of hypothesis tests of θ.

Given Example 1A as a frame of reference, I will now present two basic principles. The first principle is the *Weak Conditionality Principle* (*WCP*) and it involves the notion of a mixed experiment. A *mixed experiment* is simply a randomly drawn experiment from a set of two or more experiments. In the context of the two experiments described in Example 1A for flipping a coin, define the mixed experiment E_* whereby $J = 1$ or $J = 2$ is observed, each having probability .5 independently of θ, Y_1 or Y_2 and experiment E_J is then performed. For example, suppose a "fair" six-sided die is rolled and if an even number is obtained, then E_1 is performed, and if an odd number is obtained, then E_2 is performed. Then according to the WCP when confronted with a mixed experiment in which one of two possible experiments involving the same parameter θ can be performed, and when it is known that one experiment—say the first—was actually performed, then the entire evidence about θ provided by the mixed experiment is given by the evidence provided by the first experiment. In other words, the second experiment, which could have been performed but was not, provides no evidence regarding θ. Formally, the WCP can be stated concisely as follows.

WCP. Suppose there are two experiments $E_1 = \{Y_1, \theta, f_1(y_1|\theta)\}$ and $E_2 = \{Y_2, \theta, f_2(y_2|\theta)\}$ providing evidence about the same parameter θ. Consider the mixed experiment $E_* = \{Y_*, \theta, f_*(j, y_j|\theta)\}$ where $Y_* \equiv (J, Y_J)$ and $f_*(j, y_j|\theta) = .5f_j(y_j|\theta)$. Then $Ev[E_*, (j, y_j)] = Ev(E_j, y_j]$.

The second principle is the *Weak Sufficiency Principle* (*WSP*), and it involves the notion of a sufficient statistic. Recall that a sufficient statistic for θ is a known function of the data, say $W = W(Z_1, Z_2, \ldots, Z_T)$, such that the conditional distribution of $Z_t(t = 1, 2, \ldots, T)$ given W does *not* depend on θ. In other words W is sufficient for θ if all the information about θ contained in the data is obtained from consideration of θ alone. The WSP states that when a sufficient statistic exists, then two samples which yield the same observed value for the sufficient statistic provide the same evidence regarding θ. Formally, the *WSP* can be stated as follows.

WSP. Consider an experiment $E = \{Z, \theta, f(z|\theta)\}$ and suppose $W(Z)$ is a sufficient statistic for θ. Let z^1 and z^2 be two realizations of E and suppose $W(z^1) = W(z^2)$. Then $Ev[E, z^1] = Ev[E, z^2]$.

The WCP and the WSP seem intuitively plausible to many researchers, including myself. Their acceptance, however, logically necessitates (as shown by Birnbaum, 1962) acceptance of a third principle which proves more controversial. This third principle is the *Likelihood Principle* (*LP*) and it states that two experiments providing evidence about the same parameter θ which give rise to data realizations yielding likelihoods which are proportional, must provide the same evidence regarding θ.[5] Formally, the *LP* can be stated as follows.

LP. Consider two experiments $E_1 = \{Y_1, \theta, f_1(y_1|\theta)\}$ and $E_2 = \{Y_2, \theta, f_2(y_2|\theta)\}$ involving the same parameter θ. Suppose that for particular realizations y_1 and y_2 of the data, $L_1(\theta; y_1) = cL_2(\theta; y_2)$ for some constant c not depending on θ. Then $Ev[E_1, y_1] = Ev[E_2, y_2]$.

Note that the *LP* does not address the choice of the likelihood function (a point addressed later); rather it conditions upon a specific choice. Neither does the LP itself provide a recipe for presenting "evidence;" rather the *L* serves as a litmus test for discarding some procedures for providing "evidence." A case in point is now provided.

Example 1B. Suppose $T_1 = 12$ and $m = 3$ in Example 1A. If $y_1 = y_2 = 9$ are observed, then

$$L_1(\theta; y_1) = \frac{12!}{9!3!}\left[\theta^9(1-\theta)^3\right] = 220\left[\theta^9(1-\theta)^3\right],$$

$$L_2(\theta; y_2) = \frac{11!}{2!9!}\left[\theta^9(1-\theta)^3\right] = 55\left[\theta^9(1-\theta)^3\right].$$

Since $L_1(\theta; y_1) = cL_2(\theta; y_2)$, where $c = 4$, the *LP* is applicable, and hence, the evidence regarding θ should be the same for both experiments. To see that this prescription is not always followed, consider testing the null hypothesis H_0: $\theta = .5$ (i.e., the coin is fair) versus the alternative hypothesis H_1: $\theta > .5$ at the .05 level of significance. The "prob-values" of H_0 under the two experiments are easily seen to be $\text{Prob}(Y_1 > 9|H_0) = .0730$ and $\text{Prob}(Y_2 > 9|H_0) = .0337$. Therefore, conventional frequentist-based hypothesis testing procedures imply rejection of H_0 under E_2 but not under E_1, contradicting the LP prescription that both experiments convey the same evidence regarding θ.

---

[5] Needless to say, as with most matters of deep philosophical importance, there are dissenting views regarding the *LP* (e.g., see Berger and Wolpert, 1984; Berger, 1985b and the discussions therein). Most of the dissent regards precisely what to include in θ and in "evidence." Interestingly, some of the dissent comes from Bayesians (see Hill, 1987 and Lane, 1984), but for the present purposes, such relatively minor quibbles need not be discussed.

The case discussed in Example 1B is not atypical: it is the rule, rather than the exception, that frequentist-based procedures of providing evidence violate the LP. The main culprit behind such violations is the frequentist's preoccupation with the entire sample space in defining procedures, rather than its realized counterpart: the observed data. For example, despite having proportional realized likelihoods, the two experiments in Example 1B have different sample spaces (namely, $\{0, 1, \ldots, T_1\}$ in the case of E_1 and $\{m, m + 1, \ldots\}$ in the case of E_2), and thus different procedurally-based summaries of evidence are provided. To show that such results are no fluke, consider following example drawn from Poirier (1986a).

Example 2. Suppose E_1 corresponds to data generated according to the standard linear normal regression model, and suppose E_2 corresponds to the familiar tobit model in which values of the dependent variable less than or equal to zero are "censored" and reported simply as zero. In general these two experiments give rise to different maximum likelihood (*ML*) estimat*ors*: the ordinary least squares (*OLS*) estimator for E_1 and the tobit estimator for E_2. Now consider a special case: a sample of *non-censored* observations generated by E_2. The corresponding realized likelihood for E_2 is the same as if this sample was generated according to E_1. Hence, the *ML* estim*ate* for such a sample is the same under E_1 and E_2. However, since the two underlying estimat*ors* are different, inferences regarding the unknown parameters may differ under E_1 and E_2, and this violates the *LP*. The crucial point here is that the sampling properties of estimat*ors*, which are the basis for drawing inferences for a frequentist, are based on all possible samples (i.e., the sample space), whereas an estim*ate* is based only on the observed sample. The sample spaces corresponding to E_1 and E_2 are quite distinct.

Finally, let me turn to one of the most striking theoretical implications of the LP. The LP implies that stopping rules for data collection which depend on the actual data collected, but not θ, are irrelevant for purposes of determining the evidence on θ provided by the data.

For example, in terms of Examples 1A and 1B, it should not matter for purposes of inference regarding θ whether the researcher had first decided to observe the number of successes in twelve trials, or whether the researcher decided to conduct as many trials as necessary in order to obtain three failures. For many researchers this result at first sounds like heresy because it appears to imply that a researcher can choose to sample until reaching a foregone conclusion. As Savage (1962) so aptly put it:

> I learned the stopping rule principle from Professor Barnard... Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right.

Frequentists sometimes respond that such a principle throws out all the rules. If the stopping rule is irrelevant, they say, why can't a researcher sample until reaching a foregone conclusion? Indeed it is well known (e.g., Cornfield, 1966) that if the sampling theorist using conventional (frequentist) hypothesis testing procedures can choose to stop sampling based on whatever data has been observed; then a null hypothesis can almost surely be rejected.

The *LP* proponent's response is that there is nothing inherently wrong with the frequentist's intuition: sampling to a foregone conclusion is indeed disturbing. But the *LP* does *not* imply such a conclusion. The *LP* is not prescriptive in articulating how "evidence" is determined, but rather only that in two experimental settings the evidence on θ provided by the realized data should be the same whenever the realized likelihoods are proportional. The frequentist intuition should be interpreted not against the *LP*, but rather with the manner in which testing is actually carried out. Some methods of testing, e.g., conventional Bayesian posterior odds analysis, both adhere to the *LP* and preclude sampling to a foregone conclusion (see Cornfield, 1966). Conventional frequentist based hypothesis testing procedures, however, permit sampling to a foregone conclusion when data dependent sampling rules are permitted. The "culprit" in this latter case is the conventional fixed level of significance independent of sample size.

Importance

The preceding discussion focused on theoretical implications of the *LP* and how the latter's prescriptions can conflict with standard practice. Here I propose to summarize the major thrust of the *LP* in terms of a simple *pragmatic principle of model building* (*PPMB*) that the reader can easily grasp. It is one of five such principles proposed in this essay.

> *PPMB* 1. Given the likelihood, inferences regarding its unknown parameters should be made conditional on the observed data rather than averaged over all possible data sets as required by conventional frequentist procedures.

In short the message is "condition on the observed data." PPMB 1 directs attention toward *final*, as opposed to *initial*, precision.[6] Final precision is concerned with the accuracy of an inference given the observed data, whereas initial precision is concerned with accuracy averaged over the sample space of all possible data. The former concerns the precision of an estimate whereas the latter is concerned with how precisely a procedure estimates a relationship. Frequentists couch their arguments in terms of initial precision and argue that statistical procedures should be judged based on their long-run performance. While, *ceteris paribus*, such long-run performance may be desirable, proponents of final precision argue that the best way to achieve it is through procedures that produce estimates that are judged satisfactory on the basis of final precision for the observed samples along the way.

---

[6] See Barnett (1982, pp. 181–183) or Oakes (1986, pp. 113–114) for highly readable accounts.

The following example illustrates in a very simple manner just how different are these two concepts of precision.[7]

Example 3. Consider a Bernoulli random variable Z which equals either $\theta - 1$ or $\theta + 1$, each with probability $1/2$, given the unknown parameter θ. Let Z_1, Z_2 comprise a random sample of size two drawn from this distribution. Then a 75 percent frequentist confidence set $C(Z_1, Z_2)$ for θ is given by the single random point.

$$C(Z_1, Z_2) = \begin{bmatrix} Z_1 - 1 & \text{if } Z_1 = Z_2 \\ (Z_1 + Z_2)/2 & \text{if } Z_1 \neq Z_2 \end{bmatrix}.$$

In other words, upon repeated sampling, sets constructed using this equation will contain θ approximately 75 percent of the time. Note, however, the realized values z_1, z_2 are such that if $z_1 \neq z_2$, then it is *absolutely certain* that $\theta = (z_1 + z_2)/2$. When $z_1 = z_2$ it is equally uncertain whether $\theta = z_1 - 1$ or $\theta = z_1 + 1$ (assuming no prior knowledge about θ). Therefore, the ex ante 75 percent confidence set $C(Z_1, Z_2)$ is really an averaging of two ex post sets: one at 100 percent and one at 50 percent. If a researcher is fortunate enough to get $z_1 \neq z_2$ (it happens half the time) so that θ is known with certainty, then what sense is there to reporting the 75 percent confidence set?

The distinction between initial and final precision is seldom made in the econometrics literature. I find this surprising since econometricians commonly encounter this distinction when teaching their students the "correct" frequentist interpretation of a confidence interval. Students usually interpret a 95 percent confidence interval as saying that the probability the calculated interval contains the unknown parameter is .95, and they find it perplexing when admonished that instead their "confidence" lies in the performance of the *procedure* used to construct the interval. I believe this teaching experience indicates that final precision is a "natural" way to interpret the results of a statistical analysis. Initial precision may have a role for some purposes (e.g., experimental design, which by its very nature is a pre-data activity), but inference is not one of them. Indeed few statisticians favor a completely unconditional analysis.[8] Rather many have called for some sort of partial conditioning to rescue the frequentist paradigm from the embarassment of Examples 1–3.[9] For example, conditioning on *ancillary statistics* (i.e., statistics whose distributions are independent of unknown parameters) is often advocated.[10] While conditioning on ancillaries is often a substantial improvement over unconditional procedures, it can also lead to problems due to nonuniqueness and even nonexistence. Indeed, once one realizes the need for some conditioning, it is difficult not to go all the way, adhere to PPMB 1, and condition on all the data.

[7] This example as well as others are discussed in Berger and Wolpert (1984, Chapter 2).

[8] See Kiefer (1982) for further discussion.

[9] See Berger and Wolpert (1984, Chapter 2) and Jaynes (1976) for more examples.

[10] This idea is related to the discussion of *weak exogeneity* by Engle, Hendry and Richard (1983).

At this point I will end our discussion of the Likelihood Principle. It has, I believe, great importance for model builders. For many of the activities of model builders, the Likelihood Principle renders the frequentist view irrelevant. The Likelihood Principle "traps" the model builder into thinking conditionally, but it does not provide a full recipe for such thinking. I believe that once the Likelihood Principle is accepted, the Bayesian pill becomes easier for most researchers to swallow. There still remains the perplexing question as to what is the likelihood, but I will address that later.

The Role of "Subjectivity" in "Objective" Science

Before discussing the subjectivist view of models it is necessary to defuse some prejudice by addressing a touchy subject: the role of subjective beliefs in science. My patience, like that of many, quickly becomes strained by this issue, and I will proceed as concisely as possible.

First, I will not dwell on the inevitability of the subjectivist approach based on axiomatic developments of individual behavior in a world of uncertainty (e.g., see Fishburn, 1981) since while such developments have a long history, they seem to have won few converts. I am puzzled by this fact since econometricians of the frequentist persuasion appear to take great pride in appealing to such axiomatic bases to justify the behavior of agents in their theoretical economic models.[11] As I noted earlier, I can only encourage such colleagues to be as demanding in analyzing their own behavior as they are of the actors in the parables they espouse.

Secondly, I will not dwell on whether prior beliefs exist since I believe they undeniably do. I never cease to be amazed at non-Bayesians who question the existence of subjective prior beliefs out of one side of their mouths, and who out of the other side point out how somebody else's regression coefficient is of the "wrong sign." Nor do I ever cease to be amazed at how the same researchers impose exact identifying restrictions in simultaneous equations models.[12] That prior beliefs exist seems to me to be unquestionable and not worthy of further discussion.

Thirdly, I believe the question of whether prior beliefs *should* play a role in science is likewise a red herring. Prior beliefs manifest themselves in all econometric investigations in the form of the elusive "wisdom" that senior masters of our profession bring to bear on empirical investigations. The question of whether the subjectivist approach is legitimate within "objective" science is a play on words.[13] As I have been arguing, a subjectivist focuses attention on *observables*. What could be more "objective"?

[11] For example, see Boyer and Kihlstrom (1984).

[12] Furthermore, I am puzzled at their ignorance of the convincing argument of Leamer (1978, pp. 170–181) that multicollinearity is not a problem unless there exists uncertain prior information.

[13] See Zellner (1985, p. 255).

Fourthly, I believe subjective prior beliefs should play a *formal* role so that it is easier to investigate their impact on the results of the analysis. Bayesians must live with such honesty whereas those who introduce such beliefs informally need not.

Fifthly, I do not wish to give the impression that it is easy to articulate prior beliefs in the form of a fully specified prior distribution. As Hill (1986, pp. 223–225) notes, it is important always to keep in mind the "tentativeness" of any prior distribution. Priors are necessarily more convenient than precise. The tentativeness associated with any prior is the motivation for the important sensitivity or robustness analysis of Leamer (1978, 1982, 1983) or Berger (1984; 1985a, pp. 195–252).

Finally, since science is a public exercise, I think it behooves the researcher to give the reader a feel for how different priors, at least those reflecting established professional views, impact on the conclusions of an empirical exercise. As Ed Leamer puts it: "the message is in the mapping from the prior to the posterior." Unfortunately, this principle is easier to preach than to practice. I fear that the number of established views in economics may quickly swamp the number of mappings both researchers and editors may be willing to entertain.

In short, while admitting the existence, role and legitimacy of subjective prior beliefs, I do not believe a researcher can accomplish much by merely publishing the results corresponding to a single prior. This leads to my second principle.

> *PPMB* 2. The use of subjective prior beliefs must be accompanied by a local sensitivity analysis, and to the extent possible, a global sensitivity analysis.

The Subjectivist View of Models and Parameters

Representation Theorems

My first two PPMB's argue in favor of conditioning on the data and use of subjective beliefs in an explicit manner, but tempered by a sensitivity analysis. The first point, which is largely predicated on the *LP*, throws into question most frequentist procedures. The second point opens the door to the subjectivist arena, and it is now time to step inside and see what it is all about.

A crucial issue not yet discussed is the link between the observable world of data and the metaphysical world of models. The missing link has been provided by de Finetti (1974, 1975) and the numerous extensions of his so-called representation theorem. De Finetti wanted to show that under a weak set of conditions, the beliefs of individuals (that is, their subjective probability distributions) could all be represented in terms of a *common* likelihood function with possible differing priors. As in the coin-tossing example presented earlier, de Finetti focuses on Bernoulli random quantities. He assumed that the beliefs of individuals were *coherent* (as defined earlier) and possessed a second property called *exchangeability*.

Exchangeability imposes a form of "symmetry" on sequences of observables. In the context of Example 1A, exchangeability implies that any two sequences containing the same number of successes and failures must be assigned the same subjective probability. This numerical probability, however, is only restricted by the minimal rule of coherence, and so two coherent individuals whose beliefs are both individually exchangeable need not assign the same probabilities to particular sequences. A general definition is as follows.

> An individual's beliefs regarding a finite set of observable random quantities $Z_1, Z_2, \ldots, Z_n$ described by a joint probability distribution are said to be *exchangeable* if and only if the distribution is invariant under all permutations of the subscripts $1, 2, \ldots, n$. Beliefs regarding an infinite sequence are said to be exchangeable if and only if the induced beliefs are exchangeable for every finite subset.

De Finetti's theorem and its extensions imply that coherent like-minded individuals who share symmetries (like exchangeability) in their beliefs are led to common windows (called likelihoods) through which to view the world.[14] These windows are summarized in terms of mental constructs (unobservables) called *parameters* (like the probability of a "success" in Example 1). Like-minded individuals agree on the usefulness of these metaphysical summaries, but they may, generally, have different opinions (often called "priors") concerning them.

A group of researchers whose individual beliefs are coherent and exchangeable, but not necessarily identical across individuals, will agree on the likelihood, but they will in general have different priors for θ. In this case the likelihood is merely a euphemism for the common symmetries in beliefs a group of researchers hold regarding their uncertainties of observable events in the world. As Dawid (1982a, 1984) notes, the likelihood introduces an "intersubjective" element into otherwise personal analysis which facilitates communication among researchers. Indeed, the likelihood is a common window through which these researchers view the world. These symmetries place various restrictions on the joint distribution of observables sufficient to obtain a factorization involving a familiar form for the likelihood $L(\theta; z)$.

In practice researchers do not proceed literally in the fashion just described. Instead they directly announce the likelihood itself without deriving it with the aid of a representation theorem. I do not condemn such an approach, but I fear that researchers may easily forget the major thrust of de Finetti's Theorem: choice of a likelihood, like the prior, is a subjective exercise. The difference lies in the hope that a wide range of researchers will find the choice of the likelihood to be largely noncontroversial, and agree to disagree in terms of choice of the prior.

Since de Finetti's theorem relegates parameters to a metaphysical role, the reader may be wondering what activity should replace hypothesis testing for the scientificminded researcher. I suggest that the answer is "induction." In its simplest terms *induction* is concerned with learning from experience in order to make statements about future observables. According to the *predictivist paradigm*, articulately espoused by Geisser (1980, 1985), the ultimate goal of science is to predict future observables based on past observables. For predictivists a parametric model may be a means toward achieving this goal, but estimating the parameters of such models is *not* itself the goal. Rather, the goal is computation of a predictive distribution function of the form $G(z_{n+1},\ldots,z_{n+m}|z_1,\ldots,z_n)$ which is free of unknown parameters and from which probabilistic statements concerning the yet to be observed $Z_{n+1},\ldots,Z_{n+m}$ can be made based on having observed $Z_1 = z_1,\ldots,Z_n = z_n$.

---

[14] For example, if the sequence in question consists of real-valued random quantities for which coherent beliefs about $Z_1 - \bar{Z}, Z_2 - \bar{Z}, \ldots, Z_n - \bar{Z}$, where $\bar{Z} \equiv n^{-1}\Sigma_{i=1}^{n} Z_i$, are invariant for all n under orthogonal transformation, then such beliefs can equivalently be viewed *as if* the elements in the sequence are independent and identically distributed normal random variables with a prior for its mean a variance. Other generalizations can be found in Dawid (1982), Diaconis and Freedman (1980, 1986), Skyrms (1984) and Smith (1984, p. 252).

The predictive procedure implied by de Finetti's theorem can be shown to embody Bayes' Theorem; that is, a modification of the prior density for θ conditional on the observed data. The resulting posterior is then multiplied by the likelihood function for the future observations, and the parameters are marginalized out to yield the predictive density of $Z_{n+1},\ldots,Z_{n+m}$ conditional on observing $Z_1 = z_1,\ldots,Z_n = z_n$.

It should be clear from the foregoing that exchangeability is a weaker restriction than independence. If subjective beliefs on $Z_1, Z_2,\ldots$ are independent, then there can be no learning from experience. In contrast, exchangeability of beliefs on $Z_1, Z_2,\ldots$ implies an isomorphic view of the world in which there exists the fictitious random variable θ such that $Z_1, Z_2,\ldots$ are independent conditional on θ (that is, *conditionally independent*). Conditional independence is a concept of great importance (Dawid, 1979) and in the subjectivist framework one role for parameters is to induce conditional independence and the simplifications inherent in it.

Discussion

Many Bayesian researchers (e.g., Bayarri, *et al.*, 1986 or Box, 1980) argue that it is most useful to think of the *model* as the joint distribution of all observables and unobservables under consideration, or more generally, $L(\theta; z)q(\theta)$. The separation of the likelihood from the prior is a subjective exercise, and θ need have little meaning independent of the likelihood.[15] An important role for θ is to serve as a transmitter for non-data-based information (e.g., economic theory). Such parameters need not "exist" in the external world, but only in the minds of researchers.

If parameters are the creation of individual researchers, then the question arises: "Why spend so much time doing hypothesis testing and interval estimation?" Predictivist statisticians often ask this question of econometricians, and I think a serious response is warranted.

---

[15] This view underlies the criticism of the *LP* by Lane (1984, 1986). See also Dawid (1985a, pp. 109–110).

As I argued earlier, full articulation of subjective beliefs is not an easy exercise.[16] In order to get a handle on the problem it seems essential to summarize beliefs in terms of a model $L(\theta; z)q(\theta)$ is which the likelihood $L(\theta; z)$ is general enough to provoke interest within a sizeable audience of like-minded researchers who can then distinguish amongst themselves in terms of a (hopefully small dimensional) prior $q(\theta)$. In this light it is reasonable to pay some attention to the parameters θ in order to understand the marginal distribution of the observables. Similarly, examining the posterior distribution for these parameters provides a useful way of summarizing the predictive distribution for future observables. It does not diminish the importance of the predictive distribution to note that while it involves only observables, it is nonetheless often quite complicated. In areas like economics where large bodies of theory often focus on a few parameters for ease of exposition, researchers naturally find it easy to digest what they have learned from the data by looking at the posterior for these parameters. Somewhat paradoxically, a predictive distribution (especially one over many out-of-sample points) involving only observables may become more comprehensible by considering the underlying posterior distribution of unobservables (i.e., parameters).

Despite the apparent usefulness of parameters as a *means*, we can learn an important lesson from the predictivists. I believe this lesson is that one's concern for the posterior should be directed primarily toward gaining a better understanding of the predictive distribution. Prediction is a fundamental aspect of scientific activity and I think the pendulum has swung too far in econometrics towards testing of theories and model selection under the guise of "truth seeking." However noble the latter search may appear, I think it is largely a metaphysical exercise.

In short, there is a spectrum of opinion on the role of parameters. On one hand some researchers view them as important "lubricants" for organizing subjective beliefs, but on the other hand, some researchers appear to view them as "waste products" of prediction.[17] The attitude towards models and parameters that I would like the reader to come away with from this section is capsulized in a third PPMB.

> *PPMB* 3. Likelihoods and parameters are more usefully viewed as artifacts of the mind rather than intrinsic characteristics of the observable world. They can provide useful windows through which like-minded researchers can view the observable world and engage in prediction of future observables based on what has been observed in the past. To the extent that hypothesis testing and interval estimation facilitate such prediction, they have a useful role to play.

---

[16] For a pure subjectivist the existence of a subjective probability distribution for observables is not an issue since they accept the axiomatic developments that have been proposed for guaranteeing precisely such existence. While I find such developments quite acceptable, that does not mean that I think it is easy to operationalize the elicitations demanded by the axioms.

[17] See Geisser (1982, p. 32) who goes on to note: "Current pedagogy mistakes the lubricant for the seminal substance."

The Mechanics of Model Building

Overview

I wish to elaborate on the mechanics of model building from the subjectivist perspective. In order to be concrete, I will couch the discussion in terms of the window most familiar to economists, i.e., regression analysis. Before doing so in Example 4B, let me first set the stage.

Example 4A. Let $z_t \equiv [y_t, x_t']'$ be a $(K+1) \times 1$ vector with x_t itself a $K \times 1$ vector, and suppose T observations are stacked to form the $T \times (K+1)$ matrix $Z = [y, X]$. Consider a subjectivist researcher interested in predicting y_{T+1} given x_{T+1} and the T observations in Z. The predictive density of interest is denoted $p(y_{T+1}|Z, x_{T+1})$. At this point in the analysis no model is postulated; only an inductive question has been posed, namely: given Z and x_{T+1}, what are professionally interesting subjective beliefs concerning the yet to be observed y_{T+1}?

The situation described in Example 4A logically precedes the modeling exercise since it is first necessary to identify the thing to be modeled. Clearly, theoretical considerations enter into the choice of distribution of interest (e.g., through the choice of predictors in x_{T+1}).[18] At this stage of public research, the maximum size of the researcher's audience is largely determined. Characterization of the "professionally interesting subjective beliefs" to be entertained requires some simplifying assumptions and this leads to the introduction of a parametric model to capture the conceptual connection between y_{T+1} and x_{T+1}.[19] Ideally the likelihood component is deduced from commonly shared restrictions across researchers on subjective beliefs concerning sequences of observables; in reality it is simply announced. The intent, however, is the same: it is hoped that a variety of researchers find that the parametric window defined by the likelihood provides an attractive medium for viewing the observable world, with remaining differences captured by a manageable class of prior densities.

Let me now further elaborate upon Example 4A.

Example 4B. As an answer to the question posed at the end of Example 4A, suppose it is decided to consider as a model the standard linear normal regression

$$y = X\beta + u \tag{1}$$

together with an announced class Γ of priors over the mental construct $\theta \equiv [\beta', \sigma^2]'$.[20] In this case the likelihood is

$$L(\theta; Z) = (2\pi\sigma^2)^{-T/2}\exp[-(y - X\beta)'(y - X\beta)/2\sigma^2]. \tag{2}$$

---

[18] The choice of the distribution of interest is related to the usual parametric matter of "parameters of interest," which itself logically precedes discussion of matters of exogeneity. See Engle, *et al.* (1983) for details.

[19] Presumably this interest reflects belief in a "causal relationship" between x_t and y_t. I do not intend here to open the inevitable can of worms that presents itself in trying to explicate the meaning of "causality," but I direct the reader to Poirier (1987a) for a subjectivist treatment.

[20] For example, Leamer (1982) takes Γ as the class of normal-gamma priors with prescribed prior mean for β and arbitrary covariance matrix. If desired further restrictions on the latter can be imposed in order to further reduce the scope of Γ.

Over the class Γ of acceptable priors, there corresponds a class of predictive beliefs represented by densities of the form

$$(3)\quad p(y_{T+1}|Z, x_{T+1}) = \int_{R^K xR^+} (2\pi\sigma^2)^{-1/2} \exp\left[-(y_{T+1} - x'_{T+1}\beta)^2/2\sigma^2\right] p(\theta|Z, x_{T+1})\, d\theta,$$

where

$$(4)\qquad p(\theta|Z, x_{T+1}) = \frac{q(\theta)L(\theta; Z)}{\int_{R^K xR^+} q(\theta)L(\theta; Z)\, d\theta}$$

is the posterior density for θ given the prior density $q(\theta) \in \Gamma$.

From the subjectivist standpoint, the modelling exercise leading to the familiar regression (1) proceeds as follows. Thinking in terms of θ defines the $K + 1$ dimensional window (2) corresponding to (1). Given prior beliefs represented by $q(\theta) \in \Gamma$, Bayes' Theorem (4) is used to update these beliefs in light of the data to obtain posterior beliefs characterized by $p(\theta|Z, x_{T+1})$. Primary attention, however, focuses on predictive density (3) since it involves only observables once the mental construct θ is integrated-out. Customary econometric activities (e.g., estimation and hypothesis testing involving θ) in effect focus attention on posterior density (4). Since this posterior is an intermediary point along the mapping from Γ to (3), such activities can serve a useful conceptual role in a researcher's learning process. The principle goal of public service for the subjectivist, however, is providing the reader with a feel for the sensitivity of (3) with respect to different choices of priors drawn from Γ. The attention paid to posterior (4) should be proportional to its value in aiding the understanding of the sensitivity of predictive density (3).[21]

The reasoning behind the choice of window, such as likelihood (2) in Example 4B, is a matter of great importance in both subjectivist and other approaches. Unfortunately, it is also a matter that space constraints do not allow me to address in much detail.[22] I do wish, however, to make one point. A window is intended to be an abstraction. Whether the world is, say literally normal, linear, homoscedastic and temporally independent given θ as likelihood (2) suggests, is *not* the issue. What is important is whether the researcher and a sizeable group of readers are willing to act

---

[21] The attention paid to parameters is also positively associated with the extent to which they emit a "physical" interpretation as in a classic pure measurement error model (recall the problem of estimating the length of the farmer's field). In economics I believe parameter rarely have "physical" interpretations. In Dawid's (1985a) terms, these two interpretations are labelled "extrinsic" and "intrinsic." I do not dwell on this point, however, because the subjectivist approach is largely ambivalent on this matter: its major thrust is the same whether or not such an interpretation holds. The "existence" of parameters in the outside world is a metaphysical question of little importance to the operation of the subjectivist paradigm.

[22] I should add that I believe the choice of window involves an "artistic" aspect whose essence may not ever lend itself to complete formalization.

as if these properties hold in organizing their subjective beliefs over a specific sequence of future observables.[23]

Clearly the researcher cannot escape the question: how large a window should be adopted? In general small dimensional windows are conceptually attractive, but large windows hold the promise of a wider audience. At the end of the potential infinite regress of window expansion lies the infinite dimensional window of non-parametrics. Somehow most researchers tentatively propose a finite dimensional window and thereby judge large parametric extensions to be unwarranted. In this space of entertained extensions, the researcher implicitly assigns probability of near unity to a point reflecting the currently entertained window. The next two subsections are directed toward the situations in which such an assignment is, and is not, acceptable.

Model Building Given an Acceptable Window

Given the acceptability of the window, the principle task confronting the subjectivist researcher has already been identified: investigation of the sensitivity of predictive density (3) with respect to changes in the priors belonging to Γ. In contrast to the robustness analysis commonly found in frequentist statistics, the sensitivity analysis recommended here is conditional on the observed data z and is not concerned with data that could have been observed but weren't. In other words, the recommended analysis is loyal to PPMB 1.

Readers who accept the window, but are dissatisfied with Γ, need not be satisfied by such sensitivity analyses. Those who feel Γ is too small will not be impressed by finding that (3) is insensitive to changes in the prior; those who feel Γ is too large will not be impressed by finding that (3) is highly sensitive to changes in the prior if such changes involve elements of Γ judged to be uninteresting. When the window is relatively small, a fairly rich class Γ of priors can often be entertained without great computational burden; when the window is relatively large, the class Γ of permissible priors must often be narrowed accordingly. In short, choice of the permissible class of priors like the choice of a window is a complex matter that is difficult to formalize. "Good" empirical researchers do both well, "bad" empirical researchers do neither well. The day the economics profession agrees on who are the members in each group, is the day when concrete advice on both choices will emerge. Unfortunately, I fear this day remains in the distant future.

Often researchers behave *as if* the prior $q(\theta)$ is a mixture with components that assign positive probability to lower dimensional subspaces of the parameter space corresponding to smaller "panes" in the chosen window. Whether such lower dimensional priors are "reasonable" or not is a subjective point over which many Bayesians disagree; however, such priors can clearly serve as useful reference points. To the extent that such reductions enhance understanding of the predictive and posterior distributions, or facilitate sensitivity analysis, I think they are valuable. Unfortunately, often such dimensionality reduction is misdirected in its goals: it reflects a fondness for parameters and a failure to appreciate the predictivist philosophy discussed earlier. But even taking such priors literally does *not* necessarily imply model selection among lower dimensional specifications.[24] For example, from the predictivist standpoint it is difficult to rationalize the use of a prediction from a single lower dimensional prior; rather one is led to take a weighted prediction (since the predictive distribution is itself a mixture of component predictive distributions) with weights corresponding to the posterior probabilities assigned to the lower dimensional priors.

---

[23] In reaching this decision it is important for all concerned to appreciate the implication of the chosen window in determining the influence of individual observations on predictive density (3).

The customary motivation for model selection is a "truth search." While Bayesian posterior odds analysis (e.g., see Leamer, 1978) provides an attractive means for performing such a search (especially compared to frequentist procedures with fixed levels of significance; see Zellner 1984, p. 275ff), the goal seems peculiar given that this subjectivist discussion has avoided the term "true model." I think such a noble goal reflects the naive view that science is concerned with seeking literally true data generating mechanisms. In fact I think few deep philosophical questions can be answered by looking at a single data set. Rather the deep philosophical questions focus attention on the performance of theories over many data confrontations.[25]

Lest I wander too far astray in discussing "truth search," let me summarize the subjectivist alternative in a fourth PPMB.

> *PPMB* 4. Given a window (i.e., likelihood) through which to view the world, the major task facing the scientific researcher is to conduct a sensitivity analysis of the predictive distribution with respect to as wide a range of professionally interesting priors as is possible. To the extent that simplification and model selection activities facilitate such analysis and adhere to the Likelihood Principle, they have a vital role to play. Hypothesis testing truth searchers, no matter how noble they may appear, have a relatively minor role to play in such analysis.

I do not wish to give the reader the impression that my downplaying of the importance of many of the activities that take up a large part of the time of many econometricians is merely a reflection on the subjectivist perspective I have adopted here. Indeed even from the frequentist standpoint it is difficult to justify many of the common model building activities. Let me briefly explain.

From the frequentist perspective, standard model selection procedures (for example, see Judge, *et al.*, 1985, Chapter 21) all lead ultimately to *pretest estimators* (i.e., estimators conditional on the outcome of a preliminary hypothesis test). As Hill (1986) notes, viewed from the frequentist standpoint, pretesting has an admirable origin

[24] See Leamer (1978, pp. 208–214) or Kadane and Dickey (1980). It is difficult using posterior reasoning to justify ever imposing a simplification unless one views it as an approximation and/or one works with a criterion function containing an explicit reward for parsimony. Both such motivations are terribly difficult to quantify formally, and so such motivations are usually implemented informally.

[25] I have tried to lay out a framework (namely *model occurrence*) for tackling such questions, but I do not discuss it here (Poirier and Klepper, 1981; Poirier, 1986b, 1987a). Let me also note that as far as I know, this framework is the only one which can be used to generate prior probabilities for models in data sets yet to be observed. Such probabilities are vital both for posterior odds analysis and weighted predictive analysis.

since once the frequentist researcher admits ignorance regarding the "true model" describing the external world, there is little choice but to look at the data for guidance. As a result, pretesting is the *modus operandi* of applied researchers. Nonetheless this explanation does not excuse such researchers from swallowing two bitter frequentist pills that go along with such activities: (1) under conventional loss functions, pretest estimators are *inadmissible*; and (2) the sampling distributions of the resulting estimators are extremely complicated due in part to the nonlinearities and usual discontinuities (in terms of the data) introduced at each stage of testing. How frequentist researchers can ignore these two problems, both of which are inherently frequentist in nature, escapes me. Unawareness is not the issue, since such matters have even permeated econometric textbooks (e.g., Judge, *et al.*, 1985). I guess it shows that the "teflon factor" in statistics predates its use in kitchens and current government administrations.

Is the Window Acceptable?

The reader may have noted that we have gone a long way without the subjectivist view needing to discuss the idea of a "true model." One pragmatic reason for such avoidance is that "truth" is a notoriously slippery notion. For example, consider the following exchanges between Klamer (1984) and leading economists:

Klamer: "Are you after truth?"

Lucas: "Yeah. But I don't know what we mean by truth in our business."

Sargent: "Yes... the notion of truth is a tricky one... a continuous process."

Klamer: "Listening to you, to Lucas, and to the others... I cannot help wondering what economic truth is?"

Tobin: (Laughter) "That is a deep question. As far as macroeconomics is concerned, my objective has been to have models in which behavior is assumed to be rational, in which the gross facts of economic life are explained, and which may not give great forecasts in a technical sense but at least an understanding of what is happening and what is going to happen."

In his lead article in *Econometrica*, Feldstein (1982, p. 829) offered a related and especially frank observation:

> ... in practice all econometric specifications are necessarily "false" models... The applied econometrician, like the theorist, soon discovers from experience that a useful model is not one that is "true" or "realistic" but one that is parsimonious, plausible and informative.

I conclude from these remarks that either we are searching for something that we have difficulty describing, sure to be a futile exercise, or we are simply behaving in a pragmatic fashion consistent with my description of the subjectivist position. Personally, I hope it is the latter. Unfortunately, the phrase "true model" connotes so many things that I believe economists lose sight of the actual role of the true model in the frequentist paradigm. I believe this role is primarily regulative, that is, a "good" model's predictions must in some loose sense be in agreement with what we observed. "Truth" is really nothing more than a "correspondence" with the facts, and an important role for econometrics is to articulate what constitutes a satisfactory degree of correspondence.

I believe statisticians of most persuasions are sympathetic to placing such demands on their models, but I don't think the idea of a "true model" has any more operational content than this, apart from however noble it may make one feel about one's professional activities.

As I noted earlier, the subjectivist paradigm is designed to produce coherent revisions in our beliefs about future observables in light of what we observe. Within this paradigm (and within any statistical paradigm of which I am aware), there is no guarantee that our beliefs about future observables will in fact turn out to be "accurate." For most people I believe the attraction of models that fit well is the implicit belief that they will continue to do so in the future. All forecasters make this inductive leap, but of course it has no firm deductive basis. Subjectivism alone insures only that beliefs about future observables be coherent, not that they be accurate. The idea of a "right" or "wrong" prior misses the point; as long as the prior is coherent then the subjectivist position offers no further criterion for selecting among priors beyond that of an honest assessment of one's own prior beliefs. At any point in time it is possible to do an accounting based on a family of predictive distributions in order to gauge which prior yielded through its predictive density the most accurate predictions over a past of period time. But again, moving from there to good predictive performance in the future involves the bothersome inductive leap.

I do not propose to solve the problem of induction, but I do wish to distinguish between two types of predictive comparisons: relative and absolute. If interest focuses on a comparison relative to some other model, and given a world view (window or likelihood) that encompasses both models, the analysis is conceptually straightforward, say, via posterior odds analysis. For example, in Poirier (1987b) I suggest a diagnostic test for the regression framework of Example 4B with respect to departures from the assumed covariance matrix $\sigma^2 I_T$.

If interest focuses on an absolute comparison of a model's predictions with the facts, then a very different situation results. By definition no other model is available for such a comparison. The economist is asked to judge whether a particular window provides an adequate view of reality without reference to an alternative window. Box (1980, p. 387) argues that it is a matter of "ordinary human experience" that the ability to judge that a situation is unusual does not necessarily depend on the immediate availability of an alternative.[26] Statistically speaking, the issue is one of

---

[26] In contrast, Lindley (1982, p. 81) notes that all observations have something unusual about them; the key question is whether there is a reasonable alternative to explain them. See also Zellner (1984, p. 275ff).

pure significance testing, that is, testing a null hypothesis without an explicit alternative.[27] Since no "given likelihood" encompasses the question, the Likelihood Principle is essentially silent. Nonetheless, subjectivists have tackled the problem, but they are naturally led outside the subjectivist paradigm.[28]

I believe pure significance tests are always open to criticism due in part to inherent vagueness in the problem being posed. As much as I dislike the idea of testing a hypothesis without an explicit alternative, I must admit that the same pragmatism that earlier forced me to admit that often alternative hypotheses are difficult to describe fully leads me to my last PPMB.

> *PPMB* 5 (Cromwell's Rule[29]). Never assign a literal probability of unity to the window through which you choose to view the world.

In other words, I advocate always saving ε probability for the possibility that you may encounter some data that will lead you to abandon your window. The data evidence may come via a pure significance test, or more informally, through an interocular experience (one which hits you right between the eyes). At the time no formal alternative window may be at hand, but the data nonetheless motivate a total rethinking of the problem. As Smith (1984, p. 250) notes, the resulting model formulation experience lies largely outside the purview of any statistical theory. The gestalt switch associated with the formulation of a new window is probably best left as unconstrained as possible. In discussing this issue with statisticians, I have often encountered profound words of wisdom such as: "The important thing for a researcher to do when undergoing such a switch is to *think*." I can hardly disagree with such advice, but it seems about as useful as the dictum: "Buy low and sell high."

The point I wish to make here is that while window formulation may in fact largely be the result of creative thinking, the situation faced by the researcher once the new window is formulated is not one for which statistical methodology should be silent. Indeed the issue is one of prime importance, namely, *data snooping* (i.e., looking at the data before choosing the window). As Hill (1986) argues, once the new window is formed the Likelihood Principle again applies. The challenge facing the researcher is to formulate a "prior," or more accurately, a "post-data" prior. I believe such priors exist *before* the data, but that the difficulties of formulating priors cover them up. The data encounter that provokes the window switch awakens the researcher via Cromwell's Rule to rediscover this latent prior.

Often the new window is a small generalization of the old window and hence some of the earlier prior structure may carry over. For example, Leamer (1978, pp.

---

[27] Examples in econometrics are the Box-Pierce/Box-Ljung tests for serial correlation and Hal White's information matrix tests.

[28] See Box (1980) and Dawid (1982b, 1984, 1985b).

[29] The terminology used here is patterned after Lindley (1982) who noted that Cromwell advised the Church of Scotland to remember that it might be wrong. The spirit behind PPMD 5 can be traced back to many others (e.g., Sir Harold Jeffries). The careful reader may recall my earlier suggestion of applying a probability *near* unity to one's initial window.

295–306) considers the case where a researcher, after looking at the data, inserts additional regressors in the model, and convincingly argues that the researcher has already implicitly revealed a degree of structure on the new regression coefficients (namely, a substantial amount of prior mass should be near zero).

In summary, I believe economists and statisticians should always be prepared to abandon the window through which we view the world when confronted with "convincing" evidence to the contrary. While technically speaking such a switch may amount to global incoherence, I believe this is a lesser evil than its polar opposite—dogmatic attachment to one's initial window. Once a new window is selected, it is important in selecting a "prior" to draw on whatever soul-searching done previously that is also applicable to the new window. The "data snooping" problem for the Bayesian lies in articulating reasonable rules of the game for "post-data" prior formulation. However difficult this problem may appear, I believe it is not as onerous as the pretesting problem confronting frequentists.

Why Isn't Everyone an *Announced* Bayesian?

In this concluding section I wish to briefly address the above question which paraphrases a similar question asked by Efron (1986). Efron (p. 11) readily concedes "the considerable philosophical advantages of the Bayesian approach," and argues that the Bayesian (i.e., subjectivist) approach is not "objective" and often too hard. I have already argued that frequentists have no monopoly over the term "objective," and I do not wish to get into that seemingly endless fight again. Regarding "difficulty," to the extent that the Bayesian approach is harder—and I think its difficulty is often overestimated—it is because it delivers more.[30] Frequentist estimates are almost always obtainable just as easily in the Bayesian framework, but with the awareness of the unsatisfactory assumptions (e.g., the implicit priors) that underlie them. "Good" empirical research is difficult no matter how one attempts to perform it. I think many good frequentist researchers already act *as if* they agree with many of the five principles I have outlined. The perplexing question in my mind is why they don't just admit their Bayesian sympathies.

Two answers to my question are in fact provided in the following comments of Bayesian discussants to Efron (1986). I hope they are not right, but I fear that they are. Maybe my discussants will be able to offer less incriminating answers.

> An understanding of why this kind of philosophy does not hold sway continues to elude me but seems in any case to lie largely outside the purview of Efron's rather narrowly focused analysis. Maybe it has to do with... educational and institutional inertia? Or perhaps most statisticians now *are* Bayesians (when it matters!), but they do not want to spoil the fun by admitting it.
>
> Smith (1986, p. 10)

---

[30] As Lindley (1986, p. 6) notes: "... it is always harder to adhere to a strict moral code than to indulge in loose living."

> Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit he might have been wrong.
>
> Lindley (1986, p. 7)

■ *The research support of the Social Sciences and Humanities Research Council of Canada under grant No. 410-86-0319 and the Department of Statistics of the University of Warwick are gratefully acknowledged. Comments by Steven Garber, Seymour Geisser, George Judge, Charles Manski, Marlene Puffer and Arnold Zellner on an earlier draft are gratefully acknowledged.*

References

Barnett, Vic, *Comparative Statistical Inference.* 2nd ed. New York: John Wiley, 1982.

Basu, Debabrata, "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association*, June 1977, *72*, 355–366.

Bayarri, M. J., Morris H. DeGroot, and Joseph B. Kadane, "What is the Likelihood Function?", Technical Report No. 370, Department of Statistics, Carnegie-Mellon University, 1986.

Berger, James O., "The Robust Bayesian Viewpoint (with discussion)." In Kadane, Joseph B. ed., *Robustness in Bayesian Statistics.* New York: Elsevier, 1984, pp. 63–144.

Berger, James O., *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York: Springer-Verlag, 1985a.

Berger, James O., "In Defense of the Likelihood Principle. Axiomatics and Coherency (with discussion)." In Bernardo, Jose M., Morris H. DeGroot, Dennis V. Lindley and Adrian F. M. Smith, eds., *Bayesian Statistics 2.* Amsterdam: North-Holland, 1985b, pp. 33–66.

Berger, James O., "Bayesian Salesmanship." In Goel, Prem M., and Arnold Zellner, eds., *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno de Finetti.* Amsterdam: North-Holland, 1986.

Berger, James O., and Robert L. Wolpert, *The Likelihood Principle.* Hayward, CA: Institute of Mathematical Statistics, 1984.

Birnbaum, A., "On the Foundations of Statistical Inference," *Journal of the American Statistical Association*, March 1962, *57*, 269–326.

Box, George E. P., "Sampling and Bayes Inference in Scientific Modelling and Robustness (with discussion)," *Journal of the Royal Statistical Society*, Series A, Part 4, 1980, *143*, 383–430.

Boyer, Marcel, and Richard E. Kihlstrom, *Bayesian Models in Economic Theory.* Amsterdam: North-Holland, 1984.

Cornfield, Jerome, "A Bayesian Test of Some Classical Hypotheses—with Applications to Sequential Clinical Trials," *Journal of the American Statistical Association*, September 1966, *61*, 577–594.

Dawid, A. P., "Conditional Independence in Statistical Theory (with discussion)," *Journal of the Royal Statistical Society*, Series B, No. 1, 1979, *41*, 1–31.

Dawid, A. P., "Intersubjective Statistical Models." In Koch, Georgio, and Fabio Spizzichino, eds., *Exchangeability in Probability and Statistics.* Amsterdam: North-Holland, 1982a.

Dawid, A. P., "The Well-Calibrated Bayesian (with discussion)," *Journal of the American Statistical Association*, September 1982b, *77*, 605–613.

Dawid, A. P., "Statistical Theory: The Prequential Approach (with discussion)," *Journal of the Royal Statistical Society*, Series A, Part 2, 1984, *147*, 278–292.

Dawid, A. P., "Probability, Symmetry and Frequency," *British Journal of the Philosophy of Science*, 1985a, *36*, 107–128.

Dawid, A. P., "Calibration-Based Empirical Probability (with discussion)," *Annals of Statistics*, December 1985b, *13*, 1251–1285.

de Finetti, Bruno, "La Prévision: Ses Lois Logiques, ses Sources Subjectives," *Annals de l'Institut Henri Poincaré*, 1937, *7*, 1–68; English translation in Kyburg, Jr., Henry E., and Howard E. Smokler, eds., *Studies in Subjective Probability*. New York: John Wiley and Sons, 1964, pp. 93–158.

de Finetti, Bruno, *Theory of Probability*, Vol. 1. New York: John Wiley and Sons, 1974.

de Finetti, Bruno, *Theory of Probability*, Vol. 2. New York: John Wiley and Sons, 1975.

Diaconis, Persi, and David Freedman, "De Finetti's Generalizations of Exchangeability." In Jeffrey, Richard C., ed., *Studies in Inductive Logic and Probability*. Vol. II. Berkeley: University of California Press, 1980, 233–249.

Diaconis, Persi, and David Freedman, "A Dozen de Finetti-style Results in a Search of a Theory," Technical Report No. 85, December 1986, Department of Statistics, University of California at Berkeley.

Efron, Bradley, "Why Isn't Everyone a Bayesian? (with discussion)," *American Statistician*, February 1986, *40*, 1–11.

Engle, Robert F., David F. Hendry, and Jean-Francois Richard, "Exogeneity," *Econometrica*, March 1983, *51*, 277–304.

Feldstein, Martin S., "Inflation, Tax Rules and Investment: Some Econometric Evidence," *Econometrica*, July 1982, *50*, 825–862.

Fishburn, Peter C., "Subjective Expected Utility: A Review of Normative Theories," *Theory and Decision*, 1981, *13*, 139–199.

Geisser, Seymour, "A Predictivistic Primer." In Zellner, Arnold, ed., *Bayesian Analysis in Econometrics and Statistics: Essays in Honour of Harold Jeffreys*. Amsterdam: North-Holland, 1980, pp. 363–382.

Geisser, Seymour, "Observations on Graduate Programs in Statistics and Related Issues." In *Teaching of Statistics and Statistical Consulting*. New York: Academic Press, 1982, pp. 21–33.

Geisser, Seymour, "On the Prediction of Observables: A Selective Update (with discussion)." In Bernardo, José M., Morris H. DeGroot, Dennis V. Lindley and Adrian F. M. Smith, eds., *Bayesian Statistics 2*. Amsterdam: North-Holland, 1985, pp. 203–230.

Good, I. J., "46656 Varieties of Bayesians," *American Statistician*, February 1971, *25*, 62–63.

Good, I. J., "The Bayesian Influence, or How to Sweep Subjectivism Under the Carpet." In Harper, William L., and Clifford A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Vol. II. Dordrecht: D. Reidel Publishing Co., 1976, pp. 125–174.

Hill, Bruce M., "Some Subjective Considerations in the Selection of Models (with discussion)," *Econometric Reviews*, No. 1, 1986, *4*, 191–288.

Hill, Bruce M., "On the Validity of the Likelihood Principle," *American Statistician*, May 1987, *41*, 95–100.

Jaynes, Edward T., "Confidence Intervals vs Bayesian Intervals (with discussion)." In Harper, William L., and Clifford A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*. Vol. II. Dordrecht: D. Reidel Publishing Co., 1976, pp. 175–257.

Judge, George G., William E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee, *The Theory and Practice of Econometrics*. 2nd ed. New York: John Wiley and Sons, 1985.

Kadane, Joseph B. and James M. Dickey, "Bayesian Decision Theory and the Simplification of Models." In Kmenta, Jan, and James B. Ramsey, eds., *Evaluation of Econometric Models*. New York: Academic Press, 1980, pp. 245–268.

Kiefer, Jack, "Conditional Inference." In Kotz, Samuel, and Norman L. Johnson, eds., *Encyclopedia of Statistical Science*. Vol. 2. New York: John Wiley and Sons, 1982, pp. 103–109.

Klamer, Arjo, *Conversations with Economists*. Totowa, N.J.: Rowman and Allanheld, 1984.

Lane, David A., "Discussion." In Berger, James O., and Robert L. Wolpert, *The Likelihood Principle*. Haywood, CA: Institute of Mathematical Statistics, 1984, pp. 175–181.

Lane, David A., "Comment," *Econometric Reviews*, No. 2, 1986, *4*, 253–258.

Leamer, Edward E., *Specification Searches*. New York: John Wiley and Sons, 1978.

Leamer, Edward E., "Sets of Posterior Means With Bounded Variance Priors," *Econometrica*, May 1982, *50*, 725–736.

Leamer, Edward E. "Let's take the 'Con' Out of Econometrics," *American Economic Review*, 1983, *73*, 31–43.

Leamer, Edward E., "Bid-Ask Spreads for Subjective Probabilities." In Goel, Prem K., and Arnold Zellner, eds., *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland, 1986, pp. 217–232.

Lehmann, E. L., *Theory of Point Estimation*. New York: John Wiley and Sons, 1983.

Lindley, Dennis V., "The Bayesian Approach to Statistics," In Taago de Oliveria, J., and Benjamin Epstein, eds., *Some Recent Advances in Statistics*. New York: Academic Press, 1982, pp. 65–87.

Lindley, Dennis V., "Comment," *American Statistician*, February 1986, *40*, 6–7.

Oakes, Michael, *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New York: John Wiley and Sons, 1986.

Poirier, Dale J., "Contrasts in Inferences Based on Sampling Distributions and Posterior Distributions," *Econometric Theory*, August 1986a, *2*, 287.

Poirier, Dale J., "Rationalizing the Anomalous Performance of Models in Different Data Sets," Working Paper No. 8510 (revised), Department of Economics and Institute for Policy Analysis, University of Toronto, 1986b.

Poirier, Dale J., "Causal Relationships and Replicability," unpublished manuscript, January 1987a, Department of Economics, University of Toronto.

Poirier, Dale J., "Examples of Bayesian Diagnostic Tests," paper presented at the Third Valencia International Meeting on Bayesian Statistics, June 1987b, Altea, Spain.

Poirier, Dale J., and Steven Klepper, "Model Occurrence and Model Selection in Panel Data Sets," *Journal of Econometrics*, December 1981, *17*, 333–350.

Savage, Leonard J., et al., *The Foundations of Statistical Inference*. Methuen, 1962.

Sims, Chris A., "Comment," *Econometric Reviews*, No. 2 1986, *4*, 269–275.

Skyrms, Brian, *Pragmatics and Emiricism*. New Haven: Yale University Press, 1984.

Smith, Adrian F. M., "Bayesian Statistics (with discussion)," *Journal of the Royal Statistical Society*, Series A, Part 2, 1984, *147*, 245–259.

Smith, Adrian F. M., "Comment," *American Statistician*, February 1986, *40*, 10–11.

Zellner, Arnold, *Basic Issues in Econometrics*. Chicago: University of Chicago Press, 1984.

Zellner, Arnold, "Bayesian Econometrics," *Econometrica*, March 1985, *53*, 253–269.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 145-151

Comment on Poirier: The Subjective Perspective of a "Spiritual Bayesian"

John Rust

Professor Dale Poirier has preached a rousing sermon for the Bayesian cause, and in doing so has concisely exposed some of the cracks in the foundations of statistics. One of these cracks is the very interpretation of the concept of probability, a philosophical question that has never been fully resolved and has led to the schism that currently exists in statistics between the Bayesian and classical schools. The paper is valuable for econometricians because it challenges them to think more deeply about the meaning of the statistical procedures they employ. It certainly did that for me.

However a basic question was not addressed in the paper; namely, what is the purpose of statistical inference? I think the answer to this question has at least as much bearing upon whether someone adopts Bayesian or classical methodology as whether one interprets the meaning of probability in subjectivist or frequentist terms. Certainly beliefs about the proper "purpose" of statistical inference diverge every bit as much as do beliefs about the meaning of probability. Many people, especially statisticians, think the purpose is "prediction," a view Poirier sees as favoring the Bayesian school. Others, such as Ed Prescott, think the purpose is "measurement." Under this view, the parameters to be inferred or measured are real physical quantities and the only source of randomness is measurement error. Since physical quantities have an existence independent of our mental representations of them, this view leads one to avoid the subjective Bayesian approach in favor of "objective" classical methods. Still others

■ *John Rust is Associate Professor of Economics, University of Wisconsin, Madison, Wisconsin.*

(myself included), think that the purpose is to "learn." This view treats statistical inference as a process of formulating a sequence of *models* that are abstract but in some sense increasingly accurate representations of certain aspects of reality that we seek to understand. Although difficult to define precisely, learning typically involves three phases: observation and summarization of data, interpretation of the evidence, and revision of previously held beliefs. Bayesian decision theory is perhaps the most well-known formal model of learning that incorporates all three of these phases. Thus, to the extent that we use statistical inference to learn from data, it seems that rational scientists "ought" to be Bayesians.

Limitations of Bayesian Estimation Theory

As Poirier has observed, many economists are somewhat schizophrenic in their application of Bayesian decision theory. The schizophrenia is aptly summarized in a joke Thomas Sargent once told about one of his colleagues: "He models all his agents as knowing Bayesian decision theory even though he doesn't." Actually, the colleague in question is an expert in Bayesian decision theory; he simply chooses not to use the theory to govern the way he does empirical work. Many other applied economists behave this way, often unwittingly. We are perfectly willing (probably even compelled), to model all our agents as hyper-rational Bayesian decision makers, yet very few of us actually use the Bayesian theory when we do our empirical work. How can we explain the apparent hypocrisy?[1] As a general response to the points in Poirier's paper, let me list five reasons why I think many applied economists fail to use formal Bayesian methodology, even though they seem to understand the theory and use it to build their economic models.

First, as many Bayesians such as Poirier and Leamer admit, it is very difficult to quantify our subjective beliefs formally in terms of a prior probability distribution over an abstract parameter space. Although I think it is clear that all economists have a more or less definite system of initial beliefs before they analyze a given data set, the internal, intuitive representation of these beliefs does not appear to correspond well to a subjective probability distribution over an unknown "parameter vector." It may be true that we can be viewed as acting "as if" we had such a prior distribution; however, the process of attempting to elicit these beliefs in terms of a formal prior has proved to be a very difficult and often very arbitrary exercise.

---

[1] The hypocrisy highlights the distinction between the "normative" vs. "prescriptive" roles of Bayesian decision theory. The hypocritical behavior of econometricians shows that Bayesian decision theory is not a good normative "meta-theory" of how econometricians actually learn, nor is it a good prescriptive theory of how we "ought" to learn. I believe that our real learning process is a far more complex and creative form of behavior than the simple Bayesian theory admits. Economists who model their agents as Bayesian learners do so primarily because of analytical convenience and convention rather than out of a belief that agents "really" learn that way. Thus, our hypocrisy may in fact be a reflection of a sensible learning strategy known as *Occam's Razor*: start with the simplest possible model capable of explaining the data. At the level of abstraction of current models, building more detailed models of learning by economic agents is probably not called for.

Second, even if the applied economist succeeded in eliciting a personal prior and has formally conditioned on the data to derive a posterior, it is not clear why a reader of a journal article should be interested in the author's subjective prior and posterior. Poirier concedes this point, and stresses Leamer's slogan "the mapping is the message." Under this view, an empirical paper should present the relevant information to allow the reader to perform the mapping from prior to posterior, or at least conduct sensitivity analyses showing the impact of changes in the author's subjective prior on the calculated posterior. Unfortunately, there seems to be no general, convenient method for communicating this mapping or conducting general sensitivity analyses in a published journal article. If we follow the Bayesian logic to its limit, empirical papers would consist of a likelihood function, a dataset, and an exhortation to readers to find their own posteriors. At this latter extreme, the information content of published empirical work starts to evaporate; the reader is essentially forced to redo most of the empirical work himself. This defeats one of the basic functions of learning: to summarize complex data.

Third, the statistical models that are of interest to applied economists are rarely the simple textbook linear or Bernoulli models that admit convenient conjugate priors. Instead, interest has focused on increasingly complex dynamic, structural choice and equilibrium models whose functional forms are highly nonlinear in parameters, and for which even a single likelihood function evaluation can be relatively expensive to compute. For these types of models the computational cost of a formal Bayesian analysis (which would be forced to perform direct numerical or Monte Carlo integration over the entire parameter space in order to map the prior into a posterior) appears to be prohibitive. Die-hard Bayesians who insist on limiting the scope of their economic models to the simple textbook functional forms and specific priors which admit convenient posterior calculations have become increasingly susceptible to the criticism that their models are contrived, that their inferences are fragile, and that their models have limited economic content.

Fourth, Bayesian methodology is heavily dependent on the use of parametric functional forms for the likelihood function. However, if you ask most applied economists, they will admit that they are not at all confident about their choice of a particular functional form, especially when it comes to a parametric specification of unobservable variables that are typically included to obtain a nondegenerate statistical model.[2] Over the past ten years there has been substantial progress in the development of flexible, semi-parametric, and nonparametric statistical methods which in effect allow the researcher to specify infinite-dimensional parameter spaces, avoiding arbitrary choices of parametric functional form. By and large, most of this work has been done in the classical tradition. In principle, one would like to do nonparametric Bayesian analysis, with a prior over an infinite-dimensional parameter space. Poirier concedes this point in his "pragmatic principle of model building 5": never assign prior probability 1 to a particular parametric likelihood function through which you view the data. Unfortunately, there are some very serious problems confronting the development of a formal, nonparametric Bayesian decision theory. To date, researchers have encountered difficulties specifying interesting, non-degenerate priors over infinite-dimensional spaces. Even in cases where well-behaved parameter spaces and non-degenerate priors have actually been formulated, there is an additional problem that the nonparametric Bayesian approach may be *inconsistent*: the posterior distribution need not converge to a unit mass on the "true" parameter vector as the sample size tends to infinity. This is a serious problem, since it implies that the nonparametric Bayesian approach can lead us to the wrong answer; that is, it can lead to incorrect learning even given infinite amounts of time and data. It is well-known that in the case of finite-dimensional parameter spaces Bayes rule is consistent. As early as 1963, however, the Bayesian statistician Freedman produced examples of inconsistency of Bayes rule for infinite-dimensional parameter spaces.[3] In a recent article, Diaconis and Freedman (1986) present an even more disturbing example of an inconsistency in the context of a simple nonparametric model of location. Without going into detail about their results, let me quote from their conclusion (p. 14):

---

[2] A statistical model is *degenerate* if a subset of variables is an exact function of remaining variables in the model. In practice, however, we will never be able to predict economic variables perfectly so we introduce *unobservables* ϵ_t to account for the discrepancies. For example, in the case of dynamic choice models the data $\{d_t, x_t\}$ consists of observations on agents' decisions d_t and states x_t. Blackwell's Theorem implies that under general conditions the solution to the dynamic programming problem takes the form of a *decision rule* $d_t = f(x_t, t)$, for some nonstochastic function f. Thus, if the econometrician observes all of x_t, the resulting choice model is degenerate: knowledge of f would allow us to perfectly predict agents' choices. However if we include a vector ϵ_t of state variables observed only by the agent, we obtain a nondegenerate model $d_t = f(x_t, \epsilon_t, t)$. Integrating over the unobserved variables ϵ_t produces a nondegenerate conditional choice probability $P(d_t|x_t)$ that can be used as a basis for maximum likelihood estimation of the decision rule f.

> Often a statistician has prior information about a problem, but does not really have a sharply defined prior probability distribution. Many different distributions would have the right qualitative features, and a Bayesian typically chooses one on the basis of mathematical convenience. In smooth, low dimensional problems, this ought to help, and anyway cannot lead to disaster, because the data will swamp the details of the prior. Unfortunately in high-dimensional problems, arbitrary details of the prior can really matter; indeed, the prior can swamp the data no matter how much data you have. That is what our examples suggest, and that is why we advise against the mechanical use of Bayesian non-parametric techniques.

My fifth and final point is that although Bayesian decision theory is one of the best known formal models of learning, it is not the only model of learning, and moreover, it is not clear that it is a literally correct model of how human beings actually learn. Any attempt to "mechanize" our learning process through rigid adherence to a formal mathematical model of learning may end up inhibiting the learning process, especially if that model is not a correct model of how we actually learn. My colleague Charles Manski has elsewhere referred to overly rigid adherence to Bayesian techniques as a "straightjacket" that can inhibit the creativity of researchers in their attempts to learn from data. He points out that the very completeness and internal logical consistency of the Bayesian approach can limit the creative freedom of applied researchers. Classical statistical theory, on the other hand, is incomplete, and in some instances even internally inconsistent (especially with regard to the theory of hypothesis testing). But in the hands of intelligent researchers, classical statistical methods can offer a wider and more flexible array of techniques for drawing inferences from data, especially in the case of model selection. The extra freedom and creativity permitted by the very limitations and incompleteness of classical statistical theory may aid the learning process more than it hinders it.

---

[3] Freedman's (1963) example used an infinite-dimensional parameter space of the set of all probability distributions over the integers. Thus, a prior distribution in Freedman's set-up is a probability distribution μ over probability distributions on the integers. Freedman constructed an example where the true parameter generating the data is a geometric distribution with probability of success equal to 1/4. Freedman constructed a prior μ such that as the number of observations tended to infinity, the posterior converged weakly to a unit mass on a geometric distribution with probability 3/4. Freedman's prior μ was not a "quirk:" he showed that the set of prior probability distributions μ such that the posterior distributions are consistent is a set of the first category in the weak topology on probability distributions; that is, the set of consistent priors is topologically "negligible."

Limitations of Classical Estimation Theory

I do not want my criticisms of Bayesian methodology to be interpreted as acceptance or contentment with classical methodology. Quite the contrary, I feel there are major unsolved problems affecting all aspects of classical statistical theory. The most serious of these problems, I believe, are the issues of pretesting and model selection.[4] Classical distribution theory and hypothesis testing procedures do not allow one to condition on the results of previous estimations and hypothesis tests performed as part of a specification search for the "true" model. As a result, the classical distribution theory for estimators and test statistics is likely to be incorrect once we take this conditioning information into account. A closely related limitation of classical distribution theory is the failure to account for increases in model complexity as sample size increases. Standard asymptotic distribution theory assumes that the dimension of the parameter space is fixed, computing the approximate distribution of an estimator or test statistic as the sample size tends to infinity. However, researchers usually attempt to estimate the most complex, realistic model possible, using as many parameters as their sample size permits. In practice, this leads to a parameter space whose dimension is an increasing function of sample size. For example, in a study of estimated wages equations, Roger Koenker (1985) found that the number of parameters grew roughly proportional to the 1/4 power of the sample size. We need to generalize asymptotic distribution theory to account for such increases in model complexity as the sample size grows. Initial work in this direction has been undertaken by the statistician Stephen Portnoy (1985).

I believe intelligent researchers are aware of these problems and make internal, intuitive corrections to the standard distribution theory in the process of drawing conclusions from the data. Obviously it would be much better to attempt to generalize classical distribution theory to account explicitly for pretesting and specification searching. Unfortunately, this task appears to be extremely difficult. Until solutions are found, researchers are faced with a choice: either adopt the apparently complete and logically consistent Bayesian methodology, facing the attendant difficulties already described, or else abandon this approach in favor of an incomplete classical theory that relies heavily on asymptotic distribution theory and whose practical application in applied work involves some serious logical difficulties. The philosophical shortcomings of the latter is probably the biggest reason why I regard myself as a Bayesian, at least "in spirit." However I feel that at least at present, the computational limitations of the Bayesian approach are decisive. Since I am interested in learning from data, I use classical methods because they are computationally feasible for the models I am interested in estimating. However, in interpreting the results I try to make intuitive corrections to the approximate distributions of my estimators and test statistics to account for the information I have obtained through my specification searches. You might say that I act "as if" I were a Bayesian even though I employ classical statistical techniques. I can only hope that intelligent researchers using classical methods in this way will be able to learn correctly from the data.

---

[4] It is telling that to date the most illuminating analysis of these problems is from the Bayesian perspective, in Edward Leamer's book *Specification Searches*.

Are there any guarantees that one can learn correctly using the incomplete and sometimes logically inconsistent classical theory? Recent work on formal models of learning behavior has begun to provide some answers. Bray and Savin (1986) studied the behavior of Bayesian and classical agents in a simple stochastic cobweb model. Each period agents update their estimates of a simple linear regression of the market clearing price on a vector of exogenous variables in order to determine the output quantities to supply to the market. Feedback from this learning behavior produces a linear time-varying coefficients model for the actual market price. The model does not reflect optimal learning behavior because agents operate under the continually falsified assumption that the law of motion for the market price is time invariant. However, Bray and Savin demonstrate that under a stability condition this simple learning process will converge to the unique time-invariant rational expectations equilibrium. Marcet and Sargent (1986) have extended this result to a significantly larger class of environments using a result of Ljung (1977) that relates convergence to the stability of a certain differential equation. This seems to indicate that even unthinking automatons can use classical least-squares methodology to learn the correct model despite the fact that they use misspecified models. If we allow agents to use more sophisticated classical statistical methods including hypothesis testing and time-varying parameter models, agents should be able to learn even faster, a conjecture that Bray and Savin are currently studying. While it is somewhat reassuring to know that relatively simple statistical procedures lead to correct learning in specific environments, at the present time we do not know the class of models for which this kind of stability obtains.

I think the biggest learning problem facing the applied researcher is the process of deciding whether or not a given model represents observed data well, and discovering a model that "best fits" the observed data subject to constraints of computation and identification. Viewed in this way, many practical estimation problems are essentially "specification searches" across a more or less well-defined family of alternative models. Although applied researchers might use flexible statistical techniques searching over a sequence of finitely-parameterized models, ultimately I think we are just using these parametric specifications as stepping stones for solving the real problem, which is best regarded as a nonparametric estimation over the infinite-demensional space of all models. If you accept this viewpoint, then I think even the Bayesian methodology has some serious problems of incompleteness and inconsistency. The incompleteness arises because the traditional Bayesian theory requires the researcher to specify a parametric likelihood function (that is, a model), but the theory does not tell us how to choose the likelihood function or how to modify it in light of the observed data. In practice, this question is precisely the one of interest. If a researcher attempts to make the parameter space explicitly infinite-dimensional and conduct a formal search over alternative models, one immediately encounters the problem of inconsistency of the Bayesian methodology that Freedman and others have discovered. Thus, neither the classical nor Bayesian approaches are adequate formal models of learning processes.[5]

References

Bray, M. M., and N. E. Savin, "Rational Expectations Equilibria, Learning, and Model Specification," *Econometrica*, 1986, *54*, 5, 1129–1160.

Diaconis, P., and D. Freedman, "On the Consistency of Bayes Estimates," *Annals of Statistics*, 1986, *14*, 1, 1–26.

Freedman, D., "On the Asymptotoic Behavior of Bayes Estimates in the Discrete Case I," *Annals of Mathematical Statistics*, 1963, *34*, 1386–1403.

Koenker, R., "Asymptotic Theory and Econometric Practice," Univ. of Illinois at Urbana-Champaign, manuscript, 1985.

Leamer, E., *Specification Searches*. New York: Wiley, 1976.

Ljung, L., "Analysis of Recursive Stochastic Algorithms," *IEEE Transactions on Automatic Control*, 1977, AC-22-4, 551–575.

Marcet, A., and T. J. Sargent, "Convergence of Least Squares Learning Mechanisms in Self Referential Linear Stochastic Models," Carnegie-Mellon Univ. manuscript, 1986.

Portnoy, S., "Asymptotic Behavior of M-estimators of p Regression Parameters When p^2/n is Large II: Normal Approximation," *Annals of Statistics*, 1985, *13*, 4, 1403–1417.

---

[5] My discussion suggests the need for a more accurate "meta-theory" of the econometric learning process. However in analogy to the celebrated "Gödel Incompleteness Theorem" construction of a complete meta-theory may not be logically possible due to inherent problems of self-reference. Gödel's arguments showed that any attempt to "mechanize" the deductive learning process is bound to fail: there will always exist true statements that cannot be formally deduced from any given set of axioms using a fixed set of rules of inference. Similarly, it may be the case that any attempt to formalize the inductive learning process is also bound to fail. If no antecedent limits can be placed on the inventiveness of mathematicians in devising new rules of proof, it seems even less likely that we could construct an econometric meta-theory that delimits the creativity of econometricians in learning from data.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 153–158

Comment on Poirier: Dogma or Doubt?

Adrian Pagan

For at least this century, statisticians have argued over the relative merits of the subjective and objective views of probability. Frequently the debate has been acrimonious—witness the interchange between Copas and Smith (Smith, 1984, p. 258) at the 150th Anniversary meeting of the Royal Statistical Society. The upshot of all this controversy has been that only about five to ten percent of statistics is done from the Bayesian perspective (de Groot, 1986). Clearly, worship in the Temple of Bayes is not widespread even among statisticians. Econometricians seem no more disposed towards Bayesian analysis; Poirier's (1986) own count of the journals suggests that, excluding the journal he edited, the amount of work done in econometrics explicitly in the Bayesian mode is closer to the lower than upper limit cited above.

So attachment to the Bayesian religion is a bit low. As all good pastors know, that means it is time to bring in a crusader, who will fire the faithful up and bring forth "decisions for Bayes" from among the heathens. The last two decades have witnessed a number of revivals. Zellner (1971) began it all, to be followed by the crypto-Bayesian ridge regression[1] people, the "renegade Bayesian" Ed Leamer (1983), and finally the newest seer, Poirier. Leamer and Poirier are particularly interesting crusaders. Like advertising men, they think that the failure of past crusades is largely attributable to a poor campaign, and that a new slogan would appeal to those who want to leave their old faith but are afraid to make the jump. Leamer packaged Bayes up as

---

[1] The technique of ridge regression has been interpreted as a Bayesian estimator where the investigator's subjective prior is for all the coefficients to have a zero mean; for example, see Leamer (1978). The estimator then "shrinks" the regression towards zero.

■ *Adrian Pagan is Professor of Economics, University of Rochester, Rochester, New York.*

"extreme bounds analysis," "contract curve analysis," and so on in his SEARCH program;[2] all these had a familiar and reassuring ring to economists.

Poirier goes at it a different way, appealing more to the latent statistician that he hopes is lurking behind every economist. Like a good fundamentalist preacher he demands, "Do you believe in the Likelihood Principle?" As with most doctrines, we give it a quick glance and reply, "Amen, brother." This induces the sermon. From there on we are trapped and led inexorably, via the creed of the likelihood principle (LP), to the one true religion of Bayesianism. Any who have read C. S. Lewis's account of his conversion to Christianity in *Surprised By Joy* will recognize the process.

Before I begin my critique let me compliment the preacher on his sermon. I doubt if one could find a more engaging, beguiling or coherent account of the strengths of following the Bayesian methodology. Moreover, it is very well organized and contains references rarely read by those econometricians who do not stray too far into the debates of statistics. It would be a rare econometrician who would not learn something from reading it and it would have to be made essential reading for any course that ventured to discuss methodological issues.

My job here, however, is neither just to praise Caesar nor can I hope to bury him. If outstanding statisticians have failed to come to any agreement about these matters after a century of argument, the probability that I will have anything of significance to add to the debate is so close to zero as to be indistinguishable from it. What follows then is certainly not a deep analysis of the issues raised by the paper. If anything it is a personal testimony concerning what bothers me about the dogmas of the recommended religion.

First, let us begin with the LP and example number 1B. Suppose I wanted to test H_0: $\theta = .5$ vs H_1: $\theta > .5$. To do that one constructs an index and compares its value using the data with the value it would be expected to attain if θ were .5. Many such indices are used in econometrics. Frequently they are based upon the score function (the first derivative of the log likelihood with respect to θ), $d \log L/d\theta$, as this entity will, on average, be zero when evaluated at $\theta = .5$. Now the likelihood principle says that, given two likelihoods, L_1 and L_2, that are scalar multiples of each other—that is, $L_1 = cL_2$—then the "evidence" in the two experiments corresponding to the two likelihoods is the same. In such a situation $\log L_1 = \log c + \log L_2$, making the scores $d \log L_1/d\theta$ and $d \log L_2/d\theta$ equal.

I like to think of hypothesis testing as involving two distinct stages. In the first, an index such as described above is computed and I find it natural to refer to this as the *evidence*. In the second, a *decision* is made about the hypothesis based upon this evidence. Because Poirier's example 1B features scores, and hence indexes, that are numerically identical owing to proportionality of likelihoods, I would conclude that the evidence about H_0 is the same in each experiment. But to make a *decision* about H_0 it is necessary to consider the alternative types of data that might have been generated, which means by reference to the distribution of the random variable responsible for generating the data under H_0. Because of this, different decisions in the context of Example 1B would be made depending upon whether the H_0 data was believed to have been generated by a binomial or a negative binomial density.

---

[2] Leamer (1978) describes a program, SEARCH, to implement the ideas of that book, and it was used for the example in Leamer (1983).

To me this is as it should be, and I interpret Example 1B as illustrative of some semantic problems arising with the LP rather then the death-knell for frequentist statistics. In particular, the LP's use of the term evidence seems to me to be misleading. If I interpret "evidence" in the way I have done above, then the frequentist approach would be fully concordant with the LP, making me a subscriber to the latter, since the index used to assess H_0 is invariant to the experiment. But if one incorporates a decision into the term "evidence," then I clearly would not want to accept the LP dogma. This shows just how easy it is to get carried away in the fervor of a crusade.

Of course, the discrepancy arises from the LP's advocacy of conditioning upon the data. Decisions are made only with regard to the data in hand. A frequentist statistician uses the data when computing an index, but then claims that no decision can be made without thinking of the alternative ways the data might have been generated. Even then differences between the LP and frequentist approaches need not arise. If the distribution of the index is independent of the nature of the density generating the observations, conflicts such as those in Example 1B could never happen. With a large enough sample suitably constructed indexes would be approximately normal in both experiments, and so the same decision would be made about H_0. In fact, it is well known that the posterior density coincides with the asymptotic density of the MLE for many cases (Walker, 1969). Whenever one has this situation, there can be no difference between the LP and the frequentist approach.

Let me emphasize this point with Poirier's Example 2. He envisages a case where it is known that a Tobit model generates the data, but only positive observations actually occur, so that a maximum likelihood estimate is found by regression on the (positive) observations. Now a frequentist statistician would maintain that the fact that the observed sample did not contain any zeroes should not blind us to the fact that they could have occurred. Just because the sample is sufficiently small that no zeroes have been forthcoming does not mean that we should ignore the fact that our sample is a "selected" one in which only high error individuals happen to have been included. Such "sample selectivity" should be allowed for in deciding on any hypothesis. Once again, looking beyond the data on hand in order to make a correct decision seems to me to be the appropriate thing to do.

All of the above is not to deny that there can be problems with frequentist procedures. As Poirier observes, there is a tendency for researchers to speak as if they really were making decisions only with reference to the data on hand. Within the regression context of $y_t = x_t\beta + \epsilon_t$, statements such as "strong support for the null hypothesis from the data," seem to imply that it is believed that a p-value[3] tells something about $\Pr(H_0/y, x)$. In fact it is really informative only about $\Pr(H_0/x)$, since it looks beyond the particular e_t observed (that is, y_t) to the other possible e_t's that might have eventuated if e_t is regarded as (say) a normally distributed random variable. If one really wants the first probability, and there are cases where that may well be what is most useful, then it is to a Bayesian analysis one should go. If this is not done, Tables in Berger and Sellke (1987) show quite dramatically how misleading a p-value can be for $\Pr(H_0/y, x)$, particularly if one attaches equal probability to H_0 and H_1.

---

[3] A p-value is the probability that the computed value of the index selected to assess the hypothesis could have come from the distribution that this index would exhibit if the hypothesis were true. Most basic statistics texts discuss the concept.

There is, however, a wider issue here and that is whether we should really be interested in $\Pr(H_0/y, x)$ or $\Pr(H_0/x)$ at all. The affirmative case derives from the fact that researchers all too often act as if the questions of the economics discipline are most usefully analysed by partitioning the parameter space into a point (say θ equals zero or unity) and the remainder; for example in all the debate over policy effectiveness ($\theta = 0$) or efficient markets ($\theta = 1$). McCloskey's (1985) rubbishing of this tendency should be a salutary reminder to us that such a fixation can be quite ridiculous, and a more pertinent question to ask is what range of values of θ is likely given the observations. The provision of this information is one of the strengths of the Bayesian methodology. Recently, Andrews (1986) has made an interesting attempt to use the frequentist framework for the same purpose.

Although it takes up a lot of space the LP is really not what Poirier's paper is about. The LP is a slogan and not to be confused with the true message. Like a hymn, its role is to get us in the right frame of mind so that we can be receptive to the doctrines revealed later in the paper. Unfortunately, Poirier immediately reveals himself to be a heretic in his central doctrine (PPMB 2). That priors should be subject to some type of sensitivity analysis seems perfectly sensible to me, but is anathema to the high priests of Bayesianism—witness the fact that the principal exponent of this viewpoint, Ed Leamer, is rarely asked to the altar when "true Bayesians" meet to pay homage to their creator, and never asked to officiate there. In fact, what Poirier really gives us in this paper is what I would call "eclectic Bayesianism," and it is not all clear what *its* foundations are. For example, we are told that the prior is "more convenient than precise." Later he describes a "post-data prior" that really exists before the data, but researchers actually needed to see the data before they realized what is was that they knew beforehand... After such a tortuous rationalization, the problems that Poirier professes to see in frequentist statistics seem to be rather straightforward, and it is hard for me to see why I should be happier with his current religion rather that the older one I am used to.

Moreover, all this is running perilously close to allowing the prior to depend upon the sample (within an ϵ-neighborhood in my opinion), in which case "eclectic Bayesianism" would be condemned on the grounds that it too fails to adhere to the LP. I personally feel that Poirier's brand of Bayesianism makes a lot more sense than the pure version, particularly his recognition that it is hard to believe that economists possess the precise knowledge of a likelihood and prior demanded by the formation of a posterior. But this acknowledgement leads him to concern about prior sensitivity (PPMB 4) and the use of pure significance tests (PPMB 5), neither of which has any

Bayesian foundation. After all this sinning he can scarcely adopt the "purer than thou" attitude that he does. Once the move to be ecumenical has begun it is hard to revert to the old hard-line religion that representatives of Bayesian analysis generally demanded of its promulgators, and it makes it difficult to accept the modified creed on the basis of the same arguments as are made for its parent. "Eclectic Bayesianism" may be a good approach to empirical modeling, but it has no firmer foundations than frequentist statistics.

In fact, I have become increasingly convinced that the same sort of problems bedevil all methodologies once we attempt to put them into practice. Just take the "pre-test" problem that bothers Poirier. This is really two problems. One is how to choose a model out of a set of alternatives; the other is how to interpret the resulting statistics associated with the chosen model. The first of these issues requires a frequentist to set a threshold or critical value for the indexes discussed earlier; if this value is exceeded one model would be rejected in favor of another.

To appreciate what is involved in the second, one needs to make the distinction between $\Pr(H_0/x, M_0)$, the probability of H_0 holding conditional upon a chosen model M_0 and $\Pr(H_0/x, M_0, \ldots, M_k)$, the probability of H_0 holding if we have actually arrived at model M_0 after estimating models $M_1, \ldots, M_k$. These two probabilities, referred to as conditional and unconditional, are generally not the same. Conceptually, either can be calculated in the frequentist framework. The conditional probability is a standard item in computer packages, while the unconditional probability could be found by recourse to computer-intensive simulation methods (Veall, 1986). In practice, researchers seem to prefer to report the value of $\Pr(H_0/x, M_0)$. Provided that it is recognized that what is being reported is a probability conditional upon the chosen model, this seems satisfactory. I suspect that most readers already give the latter interpretation anyway, so that search for unconditional (across models) probabilities would be irrelevant.

How does an "eclectic Bayesian" solve these same two problems? The answer is in exactly the same way. To choose between models he has to use some rule. Posterior odds is his index, but unless some critical value is set no decision would be made. Furthermore, the $\Pr(H_0/y, x)$ associated with the posterior density is quite clearly one that is conditional upon the chosen model and makes no allowance for the fact that a series of models may have actually been analyzed. Thus the "pre-test" problem occurs as much in "eclectic Bayesianism" as in frequentist statistics, and it tends to be resolved in identical ways.

Some of my oldest friends are totally dedicated to a set of religious and political principles and everything they do stems from that base. This yields a remarkable consistency in thought and action, and an attractive position to which to aspire. But whenever I try to do the same thing I find that doubt about the central dogmas starts creeping in, and I end up with a collection of rather contradictory principles. However, I doubt that my decision making would be much improved if I could manage to be consistent with a basic set of underlying beliefs. I am convinced that much the same thing happens in econometric modeling. Some people put a premium upon consistency and seek the core set of doctrines that will always achieve that result.

It seems to be this position that motivates Poirier, even if he does hedge a bit at times. If you are one of this sort of people his position might appeal to you. If you are not, then you can probably console yourself with Blackwell's reply to de Groot's question about whether Bayesian statistics was much discussed at Berkeley (de Groot, 1986, p. 48): "Not really. It used to be discussed here but you soon discover that it's sort of like religion; that it has an appeal for some people and not for other people, and you're not going to change anybody's mind by discussing it."

References

Andrews, D. W. K., "Power in Econometric Applications," Yale Univ. mimeo, 1986.

Berger, J. O., and T. Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of *p* Values and Evidence," *Journal of the American Statistical Association*, 1987, 82, 113–139.

de Groot, M., "Conversation with David Blackwell," *Statistical Science*, 1986, 1, 40–53.

Leamer, Edward, *Specification Searches*. New York. John Wiley, 1978.

Leamer, Edward, "Let's Take the Con out of Econometrics," *American Economic Review*, 1983, 73, 31–44.

McCloskey, D., "The Loss Function has been Mis-laid: The Rhetoric of Significance Tests," *American Economic Review Papers and Proceedings*, 1985, 75, 201–205.

Poirier, D., "A Report from the Battlefront," Univ. of Toronto mimeo, 1986.

Smith, A. F. M., "Bayesian Statistics: Present Position and Potential Developments," (with discussion), *Journal of the Royal Statistical Society*, Series A, 1984, 147, 245–259.

Veall, M. R., "Inferences on the Deterrent Effect of Capital Punishment: Bootstrapping the Process of Model Selection," McMaster Univ. mimeo, 1986.

Walker, A. M., "On the Asymptotic Behavior of Posterior Distributions," *Journal of the Royal Statistical Society*, Series B, 1969, 31, 80–88.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley, 1971.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 159–166

Comment on Poirier: Operational Bayesian Methods in Econometrics

John Geweke

Dale Poirier has provided a very lucid discussion of the essential differences between the frequentist and subjectivist approaches to model building in econometrics. The most disagreeable reaction I can muster is to quibble with the last pragmatic principle of model building, but that is the least important thing I have to say here and it can wait until the end. Time is better spent on the practical implications and operational considerations of subjectivist econometrics.

My purpose in this comment is to illustrate the fact that economists can indeed act upon the agenda implied by Poirier's five pragmatic principles of model building. Many economic researchers have had a subjective view of the world for a long time. But now, the argument that Bayesian approaches do not lead to operational methods, often decisive one or two decades ago, is becoming irrelevant in rapidly increasing numbers of applications as cheap and massive desktop computing power spreads. For the foreseeable future there will be substantial returns to developing appropriate numerical methods of Bayesian inference in econometrics, and to using these methods in empirical work. Many econometric problems that are messy and intractable from a classical perspective can become elegant and straightforward from a Bayesian perspective.

I am not interested in defending this statement as an abstract proposition. Rather, I am interested in providing solutions to problems that have been stubbornly immune to treatment by classical methods. I will take up two of the scores of such problems here.

■ *John Geweke is William R. Kenan, Jr., Professor of Economics, Duke University, Durham, North Carolina.*

The Linear Regression Model

Consider first what may be the most common model in applied econometrics. Based on theoretical considerations, an investigator formulates a linear model

$$Y_t = \sum \beta_j x_{tj} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2), \qquad t = 1, \ldots, T.$$

To keep matters simple, we suppose that the model and data satisfy the conditions of the Gauss-Markov Theorem and the investigator is thereby led to estimate the coefficients by least squares. Based on the theoretical considerations, the investigator also thinks that certain of the coefficients should be nonpositive and others should be nonnegative. (The theory could be pathbreaking or atrocious, but that is not the issue here.) In most cases the investigator will evaluate the theory informally by examining the least squares estimates of the coefficients, scoring high points for estimated coefficients of the right sign that are significantly nonzero by the usual standards, fewer for those of the right sign that are insignificant, and so on, with the most points being subtracted for estimated coefficients that are of the wrong sign and significantly different from zero. An example of such a model is provided in Table 1. This evaluation of the consistency of the inequality restrictions with the data is arbitrary, but it is widely used. The first column of Table 2 shows the least squares estimates of the coefficients of that model. By the informal scoring rule the model does not appear to be a failure, but the investigator probably hopes to do better.

A typical next step would be for the investigator to engage in post-specification tinkering. By systematically estimating all 2^{10} equations that can be formed by excluding subsets of variables, retaining those that have no "wrong" signs, and then reporting the set of results with highest value of R^2, the researcher will obtain the second set of results in Table 2. This, of course is "data mining." It is common. It is also—as a little analysis will reveal—maximum likelihood subject to the inequality constraints imposed by the investigator's theory. If one accepts the principle of maximum likelihood in this situation, then the strongest criticism of the investigator is that he could have read Judge and Takayama (1966) and set up a simple quadratic programming problem rather than solve 1,024 least squares estimation problems.

The first two sets of estimates in Table 2 illustrate the inadequacy of classical theory. The least squares estimates (the first set) do not incorporate the inequality restrictions at all. The constrained maximum likelihood estimates (the second set) incorporate the inequality restrictions, but the standard errors are computed as if those coefficients set to zero were known to be zero at the outset. There is an *asymptotic* theory of inference for these estimates, because when samples are large enough we are effectively certain which constraints are satisfied and which are not. That is not the case here, and in samples large enough that these asymptotics could be invoked standard errors on all coefficients would likely be so small that formal inference would be unnecessary. This illustrates Poirier's important point that the relevant question about asymptotic theory is its adequacy to the problem at hand.

Table 1

A typical linear regression model with inequality constraints

Variable number	*Definition*	*Presumed sign*
1	Intercept	None
2	Personal income less transfer payments	+
3	Index of consumer sentiment	+
4	Unemployment rate	–
5	Index of cost of car ownership	–
6	Average miles per gallon of current model-year cars	+
7	Dummy variable for automobile strikes	–
8	Depreciation rate of the stock of cars	+
9	Average price of a new car	-
10	Stock of automobiles	-
11	Interest rate on automoble loans	–

Dependent variable: Unit sales of automobiles

Data: 60 quarterly observations, see Bails and Peppers (1982, Appendix G).

Table 2

Comparison of estimates

	Presumed Sign	*Standard OLS*		*Constrained ML*		*Bayes*	
β_1	None	– 7.704	(5.380)	– 8.033	(1.641)	- 7.535	(2.129)
β_2	+	.02666	(.00743)	.02133	(.00198)	.02477	(.00231)
β_3	+	.03962	(.01694)	.04792	(.01254)	.04211	(.01230)
β_4	–	.2471	(.1561)	.000		– .02057	(.01707)
β_5	–	–5.114	(1.885)	–4.538	(.9173)	–3.274	(.8406)
β_6	+	.08402	(.2276)	.05676	(.1392)	.1080	(.0861)
β_7	–	–.1347	(.0314)	–.1281	(.0284)	–.1329	(.0199)
β_8	+	65.39	(62.09)	33 02	(43 95)	34.16	(25.56)
β_9	–	.1356	(.8960)	.000		–.3986	(.2477)
β_{10}	–	–.07232	(.09618)	.000		– .01899	(.01425)
β_{11}	–	–.008097	(.1660)	.000	- .1474	(.09023)	

Source: Geweke (1986, Table IV).

The fundamental difficulty for classical inference is that it deduces what should be observed as sample size increases if the model is correctly specified. By contrast, the objective of the investigator is to modify his view of the world conditional on a given data set, and prior information which includes the specification of the model. This modification could involve evaluating competing theories or reforming an interpretation of historical events, as well as the formation of a predictive density emphasized by

Poirier. If classical inference assists the investigator in this endeavor, that is a happy coincidence. We should not be surprised when that is not the case.

By contrast the subjective approach is natural and can be made operational as proposed in Geweke (1986). The conventional likelihood function for the normal linear model *without* the inequality constraints on the coefficients has a well known form. It is a multivariate t density with known degrees of freedom: in the two-regressor case this leads to the ellipses of concentration often seen in elementary econometrics texts. Given a conventional diffuse prior, the posterior density for the coefficients is also multivariate t, with known degrees of freedom. Using standard computational techniques it is easy to draw a random sample of coefficients from this posterior distribution. The posterior mean of function of the coefficients is well approximated by the mean of that function of the sampled coefficients. To impose the inequality constraints we follow the same procedure but discard those coefficients that violate the constraints. The standard deviation of these drawings is a numerical approximation to the standard deviation of the coefficients in the posterior distribution, and the distribution of the drawings themselves can be used to approximate the exact posterior distribution. The posterior means and standard deviations for our illustrative problem, given a prior that precludes "wrong" signs but is otherwise diffuse, constitute the third set of estimates in Table 3. Applying this method we could also compute the median of the coefficients and thereby obtain the minimum absolute deviation estimates. The mode of the posterior density, which is the maximum likelihood estimate if the prior is locally flat, is not inherently interesting.

Variants on this method provide operational Bayesian methods in a wide variety of problems. In this example, one could evaluate the inequality constraints from the perspective of an investigator who gave no credence to any of them, whose prior was diffuse for all the coefficients. The fraction of drawings from this density that obey the sign constraints stated in Table 1 is then a numerical approximation to the posterior probability of the restrictions themselves; it is 6.4×10^{5}. Notice also that constraints

Table 3

Average conditional variance of inflation in three subperiods

	Prior A	*Prior B*	*Prior C*
Average for 49:4 to 54.2	6.228(1.280)	6.464(1.373)	6.764(1.695)
Average for 56:2 to 71:2	1.090(0.176)	1.115(0.180)	1.118(0.181)
Average for 73:2 to 79.4	1.357(0.202)	1.395(0.220)	1.415(0.233)

Source: Geweke (forthcoming, Table 3).

All variances have been scaled by 10^5. All three priors are diffuse. Under "A" the prior is inversely proportional to the unconditional variance of ϵ_t. Under "B" the prior is flat on all parameters if the ARCH process is stable (equivalently, the unconditional variance of ϵ_t is defined) and zero elsewhere. Under "C" the prior is flat on all parameters so long as conditional variance is positive and zero elsewhere. See Geweke (forthcoming a) for futher details.

can be nonlinear and one can numerically construct the exact posterior density of nonlinear functions of the parameters. It is not even necessary to be able to express constraints and functions in closed form; see Geweke, Marshall and Zarkin (1986) for an example. It is also straightforward to employ almost any prior density, not just those of a conjugate parametric form—as illustrated here by the incorporation of inequality constraints in the prior density.

A More Complex Model

In more complicated models these methods remain practical and can be used to investigate directly the questions that motivate empirical work in the first place. To take a specific example, Engle (1983) investigated the relation between the variance and the level of inflation in the U.S. through the window of the model

$$\log(P_t/P_{t-1}) = \beta_0 + \beta_1 t + \beta_2 \log(P_{t-1}/P_{t-2}) + \beta_3 \log(P_{t-2}/P_{t-3})$$

$$+\beta_4 \log(PM_t/PM_{t-1}) + \beta_5 \log(W_t/W_{t-1}) + \beta_6 \log(M_t/M_{t-1}) + \varepsilon_t;$$

$$\varepsilon_t \sim N(0, h_t) h_e = \gamma_0 + \gamma \sum_{j=1}^{8} [(9-j)/36] \varepsilon_{t-j}^2.$$

(The data are quarterly, 1949–1979, and can be obtained from the *Journal of Money, Credit and Banking*. The variables P, PM, W, and M are consumer prices, an import price index, a wage index, and the money supply, respectively.) In this model the disturbances exhibit autoregressive conditional heteroskedasticity (ARCH), allowing persistence in conditional variance h_t. The restrictions $\gamma_0 > 0$ and $\gamma_1 \geq 0$ that are sufficient for positive variance can be handled in much the same way as the inequality restrictions in the previous example, although the procedure is technically somewhat more complicated; Geweke (forthcoming *a*) provides details.

Among the purposes of the empirical work, as indicated in the introduction and discussion in Engle (1983) are to determine whether or not the conditional variance of inflation differed systematically over the three decades of the sample, and establish minimums and maximums for the conditional variance of inflation over the 30 years studied. Following the original study the three periods indicated in Table 3 were considered. Table 3 provides posterior means and standard deviations for the average conditional variance of inflation in each of these periods. These reported moments are exact to at least two significant figures. The analysis here, like the question, is *conditional* on the data, which are fixed while the parameters are random. (By contrast, frequentist means and standard deviations are inappropriate to the question. Furthermore no one knows how to compute them: derivation of their asymptotic distribution

Table 4
Maximum and minimum average annual conditional variances

Year of Maximum	*Prior A*	*Prior B*	*Prior C*
1951	.9302	.9304	.9311
1952	.0698	.0696	.0689
Minimum			
1957	.2232	.2190	.2112
1959	.3297	.3288	.3339
1960	.0743	.0754	.0812
1961	.0021	.0024	.0022
1962	.1002	.1008	.1020
1963	.0454	.0458	.9461
1965	.0051	.0044	.0044
1968	.0001	.0001	.0001
1970	.0011	.0011	.0011
1971	.0741	.0753	.0738
1972	.0017	.0017	.0016
1974	.0004	.0004	.0005
1977	.1427	.1438	.1408

Source. Geweke (forthcoming *a*, Table 5. Priors are the same as in Table 3).

is an analytical challenge not likely to be taken up soon.) The three panels of Table 3 report sensitivity to three different priors, in the spirit of PPMB 2. Note that there are no estimates of parameters in Table 3: instead the table reports directly on the concepts that motivated the study, following PPMB 3.

The question of which year exhibited maximum and minimum conditional variance in inflation is not amenable to attack by frequentist methods. It is natural for methods that condition on the data set. Computation of the probability that a given year exhibited maximum or minimum conditional variance in inflation is easy following the methods used to produce Table 3. Each time a point is drawn from the posterior distribution, we compute the mean value of the conditional variance h_t for each of the years in the sample, and then determine which year exhibited the minimum value and which the maximum. Table 4 indicates all years in which one of these probabilities exceeded 5×10^{-5}.

Strength of the Bayesian Approach

With the application of subjectivist, conditional, and/or Bayesian approaches, important questions that once might have been curious metaphysical issues for the nonspecialist became practical concerns. One of the foremost of these is, in Poirier's terms, the acceptability of the window. The notion that all models are false implies that a search for unique, true data generating processes will be endless. If all models are false, then no method, including frequentist significance testing, can sustain the conclusion that no better model exists. One attraction of significance testing is that it avoids the intellectual effort of forming an explicit alternative. But rejecting a hypothesis does not mean a better model exists or can be created.

As economists acquire more experience with explicitly subjective methods, more practical perspectives on this subtle and important issue will emerge. Poirier's PPMB 5 is a salutary curb on the true believer, but strictly speaking no probability at all need be assigned to the window; one conditions on it and that is that. The important point for applied work, not elaborated in Poirier's paper, is that window expansion, addition, and other forms of carpentry are straightforward and operational. Conventional methods for the assessment of posterior odds ratios can be rendered generically operational along the lines dicussed in the empirical examples above, as detailed in Geweke (forthcoming *b*). Whether the investigator is prepared to abandon his window or not, another investigator with another window may always render it of no real importance with a sufficiently favorable posterior odds ratio. The possibility that this will happen can never be excluded.

For the formal subjectivist the loss function that motivates the investigation may indicate whether or not the model is acceptable. A loss function can provide a measure of progress in model building by means of reduction in its posterior expectation, and potential returns to additional progress can be assessed in the same way. If the loss function has a global minimum, then one may reach a point at which there is clearly no worthwhile effort to be expended in model improvement, and at that point the model may be deemed acceptable. I suspect the behavior of practical model builders more nearly resembles this process than it does any set of rules for sequential significance testing.

There can be little question that the principles set forth in Poirier's paper solve the problem of window (or model) choice from a large universe of potential models in a more satisfactory way than any frequentist procedure. Bayesian methods provide for the construction of posterior probabilities over wide and disparate classes of models. Exploiting this fact leads to analytical tasks in the formulation of coherent priors over classes of models. The important point is that there *is* a paradigm. With recent advances in computing, these methods can be applied to many of the models economists use in their applied work.

By contrast, it is very difficult to formulate model selection as a frequentist problem. In practice, significance tests and (often artificial) nesting are the most commonly used formal procedures for exploring models. To the extent that marginal variables are included or excluded based on significance tests, these procedures as conventionally employed guarantee that asymptotic theory will never apply; the regression model discussed here is a good, and common, example. Difficulties with subjectivist window evaluation pale by comparison.

■ *Financial support for this work was provided by NSF Grant SE8605867.*

References

Bails, D. G., and L. C. Peppers, *Business Fluctuations*. Englewood Cliffs: Prentice-Hall, 1982.

Engle, R. F., "Estimates of the Variance of U.S. Inflation Based on the ARCH Model," *Journal of Money, Credit and Banking*, 1983, *15*, 286-301.

Geweke, J., "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics*, 1986, *1*, 127-141.

Geweke, J., "Exact Inference in Models with Autoregressive Conditional Heteroskedasticity," In Barnett, W., E. Berndt, and H. White, eds., *Dynamic Econometric Modeling*, Cambridge: Cambridge University Press, forthcoming a.

Geweke, J., Exact Predicitive Densities in Linear Models with ARCH Disturbances," *Journal of Econometrics*, forthcoming b.

Geweke, J., R. C. Marshall, and G. Zarkin, "Exact Inference for Continuous Time Markov Chain Models," *Review of Economic Studies*, 1986, *53*, 653-669.

Judge, G. G., and T. Takayama, "Inequality Restrictions in Regression Analysis," *Journal of the American Statistical Association*, 1966, *61*, 166-181.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 167–170

The Subjectivist Response

Dale J. Poirier

It is a rare opportunity to have one's research discussed in print by as eminent and contemplative a group of researchers as my three discussants. To each I wish to express my gratitude for his thought-provoking comments. I appreciate their praise for my proficiency as a "preacher," but I do not think this is due to any special literary skills. Rather it reflects the intellectual quality of the existing scripture upon which my "sermon" was based.

I will address the discussants' comments in an order which I believe reflects a decreasing degree of common ground in our views. Given this ranking, it should not surprise the reader that I first address the like-minded comments of John Geweke.

Geweke dispels the widespread idea (for example, see John Rust's comments) that Bayesian methods are too hard to be practical. His comments are pragmatic rather than pedantic and I believe they will strike the hearts of many readers more forcefully than my theological sermon ever could. I wish to emphasize Geweke's main point: often Bayesian methods are easier. Geweke is a "doer" and his examples illustrate his main point dramatically.

Moving on to John Rust, let me say that he disarms me of any venom in my reply when he admits to being a Bayesian "in spirit" and that he acts "as if" he is a Bayesian when doing empirical research. Rust also throws me off stride when he argues (as Adrian Pagan does) that the internal logical consistency of Bayesian methods places the researcher in an undesirable "straightjacket" which inhibits creativity. I must confess that it never crossed my mind to consider "logically inconsistent" approaches.

I do not find Rust's suggestion that "measurement" serves as a primary purpose of statistical analysis to be very convincing. I am hard-pressed to think of examples in economics where the parameters of primary concern are "real physical quantities." If I could think of such examples, however, then I would proceed in an "objective" Bayesian fashion fully consistent with the LP.

I was happy to see that Rust also views the behavior of many researchers who employ Bayesian agents, but who eschew Bayesian techniques, as "schizophrenic." Of course I am also sympathetic to the difficulty of performing any sensitivity analysis, be it with respect to a prior in a Bayesian analysis or with respect to some other aspect of the model in a frequentist analysis. Good applied researchers of any persuasion are always concerned with the "robustness" of their results.

One issue raised by Rust that warrants special attention is the increasingly fashionable topic of semi-parametrics and nonparametrics. I believe he views the results of Dianconis and Freedman to be more damaging to the subjective Bayesian position than the authors themselves do, or I may add, than do many of the authors' discussants. Readers should note that Rust's characterization of the entire matter is very much in an objectivist tradition: there is something (like a conditional mean) about which we wish to learn and about which we wish to make few assumptions. In such a setting semi-parametric and nonparametric techniques appear quite enticing. As an economist, however, Rust knows there is no such thing as a free lunch, and more than most he understands the limitations of the asymptotic results that exist in these areas. (I trust the reader appreciates that, in order to judge the quality of the asymptotic approximations to a particular finite sample setting, one must make most of the assumptions one was trying to dispense with in the first place.) What I do wish to address, however, is the importance of Dianconis and Freedman's results to the subjectivist.

Once one adopts the subjectivist position that the likelihood is merely a window through which to view the observable world, the fear in making bold assumptions is lessened. The question of whether the "true" regression is linear, say, is not necessarily a meaningful question to the subjectivist. Recall that I have argued that the subjectivist window is intended to be attractive and useful. It is chosen in the hope of capturing the interests of a wide audience who will agree to disagree over the most appropriate priors to impose on the parameters which define the window. If the subjectivist fears that the professional audience is so diverse that no finite-dimensional window will ever bring about agreement over even this initial matter, then nonparametrics becomes a possible alternative. The importance of Dianconis and Freedman's results in this setting is that there may not be a tractable prior available which can guarantee that as data evidence accumulates, any two researchers need reach agreement. In effect, there is so little structure on the problem that there is no guarantee that the data will eventually swamp prior opinion. In such instances the subjectivist must be content with inevitable persistent disagreement among researchers. Casual observation of the economics profession suggests to me that this need not signal the end of research among like-minded subgroups.

Let me go further out on a professional limb (assuming this paper leaves me any further limb upon which to journey) and predict what I see as the possible future for semi-parametrics and nonparametrics in econometrics. On the positive side, we may find that after all the sexy mathematical problems have been resolved we are left with techniques that aid in choosing simple parametric windows through which to view the world. If this turns out to be the case, then it may prove to be all worthwhile. I fear, however, that the end result may be that some of the most technically-talented econometricians around may have entertained themselves for a few decades without leaving behind much that is useful for the blue-collar empiricists.

Finally, let me turn to Adrian Pagan who appears to have enjoyed my pilgrimage down the Bayesian path, but who also remains unconvinced. I am happy that Pagan addresses the LP and he raises two points which deserve comment. Firstly, although his formulation of the score test statistic remains invariant for proportional likelihoods as in example 1, this is not the case when the parent formulation using the information matrix is used, since the computation of the latter involves taking an expectation with respect to the sampling distribution which differs for the two experiments. Secondly, his comments concerning Examples 1 and 2 never escape the frequentist assumption that one must always take into account the entire sample space, and not just the observed data. If one starts with this premise, then the LP will indeed seem peculiar. The issue I had hoped he would address is which of the underlying axioms he found disagreeable.

Pagan wisely throws me off guard by claiming that my PPMB's show that I am not a true subjectivist. I am not embarrassed by my eclectic and ecumenical approach. Indeed I welcome such labels as ammunition against what I perceive will be many readers who think my position is far too radical. Let me, however, attempt to recover my subjectivist purity.

I know of few one-prior Bayesians, certainly, none among serious empiricists. The decision making agent in Savage's axiomatic development is making personal decisions and is not assumed to be behaving in the setting of the research scholar. The alleged infrequency with which Ed Leamer receives invitations to the purists' altar are equally puzzling to me. History suggests, however, that prophets are not always treated kindly, and so his calls for sensitivity analysis may some day be more appreciated.

When implementing any philosophical position reasonable researchers may take alternative branches from the parent trunk. The Bayesian trunk to which my branch is attached, however, grows far away from any frequentist forest. If I have sinned in my admission that "priors" are sometimes better described as "post-data priors," then it is at most a venial sin of the empirical researcher, and not a "purer than thou" attitude of one who has never left the sanctuary of pure theory. Since both Geweke and Pagan appear disturbed about my PPMB 5 (Cromwell's Rule), I must indeed have caused some confusion and so allow me to try to clarify matters. I think my differences with Geweke are largely semantic; I believe assigning unitary probability to a window suggests that one would never consider expanding it. The main point of PPMB 5 is to serve as a reminder to avoid arrogance. Flagrantly poor predictive performance on an absolute scale can make it abundantly clear to the researcher that "something" needs to be changed without necessarily employing any statistical tests or knowing at first what that "something" is. While this situation shares commonalities with those typically encountered when using a pure significance test, namely, the absence of a well-specified alternative window, I do not believe it implies the need for a frequentist-based pure significance test as I believe Pagan does. Nor do I believe it suggests one should abandon the old window before a new one is chosen. The creative act of choosing a new window is something I have attempted neither to describe nor explain. I suspect it will often be data-based. Once this new window is formulated, however, one can proceed in the standard Bayesian fashion to compare it to the old window as Geweke suggests. The practical importance of PPMB 5 is that the posterior probabilities for hypotheses that a Bayesian computes only measure relative performance, relative to the window in which all are nested. To assign a probability of unity to a window is to suggest these posterior probabilities measure absolute performance.

Finally, let me address Pagan's comments on pre-testing. I do not understand the argument that one should employ frequentist techniques which eschew conditioning on the data to pick a specification, and to then do the very unnatural frequentist act of conditioning on the chosen model. I agree with Pagan that people often behave this way. The rules of this "sometimes frequentist" and "sometimes conditionalist" game, however, escape me.

Pagan's characterization of the Bayesian pre-test problem also bothers me. I believe it differs from the frequentist analog in a very important way and that the LP plays an important role in this difference. For a LP proponent, how one arrives at a hypothesis does not matter for purposes of assessing the evidence provided by the data on unknown parameters. For a frequentist, two researchers who arrive at the same hypothesis following different paths will in general draw different inferences. The Bayesian view is that how one got the hypothesis should not matter in evaluating what the data have to say about the parameters given that hypothesis. The different paths, however, may reflect different prior probabilities assigned to the hypothesis and these nondata-based inputs may be quite important for testing.

Let me close by drawing the reader's attention to Pagan's closing quote of the eminent statistician David Blackwell for whom I have the greatest admiration. Pagan, Rust and myself have all engaged in religious metaphors, and given Blackwell's quote this indeed may be appropriate. Like religion, the statistical philosophy which we bring to bear on our research problems reflects to a great deal our early training, but in the latter case we receive it, not from our parents, but from the teachers of our introductory statistics courses. In this sense, not many of us are trained as Bayesians; indeed most Bayesians are converts, and we all know the evangelical tendencies of converts. Who becomes a convert is an intriguing question. I suggest a necessary condition is the willingness to read some of the Bayesian scripture, and if I have whetted the appetite of the reader to do so, then I consider my efforts here not to have been in vain. Many of those who actually do so will find it to be a "religious experience." I conjecture, however, that all who do so will find it to be an intellectual experience.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 171–178

The Status of Master's Programs in Economics

Robert J. Thornton
Jon T. Innes

It is the purpose of this article to report the results of a recent survey which we undertook concerning the content and structure of master's programs in economics. During the fall of 1985, questionnaires were sent to the 221 United States colleges and universities listed in Peterson's *Graduate Programs in the Humanities and Social Sciences* (1985) as having graduate programs in economics. One hundred sixty-nine responses were received from master's program directors, a response rate of 76 percent, which we consider to be extremely high.

Our interest in the subject began, frankly, as a parochial one. In the process of considering revisions in our own master's program, we naturally became curious as to whether the views that we held regarding such a program were at all widespread. When we attempted to discover what other programs were doing, we found that the last comprehensive report on graduate programs in economics was undertaken nearly thirty-five years earlier. That report, "Graduate Education in Economics," was written by Howard R. Bowen and appeared as a special supplement to the *American Economic Review* in 1953.[1] Since that time, economics graduate programs have proliferated in number and, as we strongly suspected, have undergone substantial changes in content and structure as well. The need for a new survey was apparent.[2]

---

[1] A more recent study was undertaken by Jack Sterrett and Saul Barr (1983). While their study generated useful information, it was far less comprehensive than the Bowen report. Moreover, their survey may also have unintentionally included information on MBA programs which offer graduate economics courses (Barr and Young, 1984, p. 4).

[2] The need for a new survey, not just of master's programs but of graduate programs in economics generally, is gradually being recognized by others in the profession. Most recently, the Committee on Economic Education of the American Economic Association expressed interest in exploring the topic of

■ *Robert J. Thornton is Chairman, Department of Economics, and Jon T. Innes is Professor of Economics, Lehigh University, Bethlehem, Pennsylvania.*

Even so, our initial purpose in undertaking the survey was simply to describe existing programs. We were, quite probably, willing to sacrifice our own vision of a proper master's program if conformity with generally accepted standards required doing so. But a perusal of the Bowen report changed our thinking. Bowen not only reported the results of his own survey, but assessed them as well. His vision of what a master's program in economics ought to be was clearly articulated, and his report was filled with recommendations for change consistent with that vision. His conclusion, that the master's degree "should be reinstated as one signifying a substantial accomplishment in economics and a level of professional competence as an economist" (p. 59), seemed to state exactly what we wanted to accomplish with our own restructuring. But our survey results suggested to us that master's degree programs have moved even further away from this ideal in the last thirty-five years. Thus our findings proved to be not so much prescriptive for our planned restructuring process as they were indicative of the deterioration of the degree.

Undoubtedly, several different models of master's programs could be constructed in accordance with Bowen's recommendation quoted above. We think that such models would still have much in common, however. For example, we believe that they would require that candidates beginning the program possess a strong initial grasp of undergraduate economics. They would also require eventual mastery of a substantial common body of knowledge beyond that expected of an undergraduate economics major. They would convey at least a moderate level of research proficiency. They would also be based on the idea that degree recipients would be certified, and could seek employment, as professional economists. It is with respect to these elemental aspects that we have attempted to judge the composite picture of the current master's degree in economics that emerges from our survey results.

Like Bowen, we do not profess to speak for anyone but ourselves in presenting our view of what a good master's program should be. Nevertheless, the very satisfying response rate to our questionnaire, and the fact that over one-fourth of the respondents added a written request for copies of our survey results, are indicative of a high level of interest in the subject. It would be surprising indeed if that interest were merely a reflection of widespread satisfaction with the current state of affairs.

Program Numbers and Admissions Prerequisites

The number of schools which offer master's degrees in economics is now approximately double what it was 35 years ago. Bowen noted that in 1951 only 135

---

graduate education. The committee submitted a proposal for funding a Symposium on Graduate Education with the papers to be published in the *Journal of Economic Education*, but as yet the proposal has not been funded ("Report of the Committee on Economic Education," *American Economic Review*, May, 1986, *76*, No. 2, 458).

institutions in the United States and Canada combined offered graduate degrees (master's or master's and Ph.D.) compared to the 221 institutions in the United States alone which offer such programs at the present time.

The great majority of institutions that offer master's degrees in economics give the M.A. degree. Of the schools responding to our survey, 87 percent offer the M.A. degree and 25 percent offer the M.S. (Twelve percent offer both degrees.) We did not seek to ascertain the difference between the requirements of the M.A. and M.S. programs, but we suspect that they are either minor or nonexistent.

We inquired in our questionnaire as to the types of undergraduate backgrounds —major, minor, or specific courses—which were required for admission to master's programs in economics. What we discovered was surprising.

Only 2 percent of those institutions with master's programs indicated that they required an undergraduate major in economics for admission into the M.A. or M.S. programs. Only 5 percent of these institutions required at least a minor in economics. Another 36 percent of institutions indicated certain course prerequisites. In most cases, intermediate level courses in microeconomics and macroeconomics were specified, and many programs also stipulated courses in statistics and calculus. The majority of responding institutions (53 percent) indicated that there were *no* specific undergraduate major or minor prerequisites for admission into the program.

The fact that a majority of master's program directors indicated that there were no specific prerequisites with respect to undergraduate major or minor fields or even with respect to specific course work is profoundly disturbing. Of course, the fact that it is possible to gain admission to graduate programs in economics with little or no undergraduate economics preparation is not necessarily tantamount to saying that the level of graduate instruction is consequently lower than it would otherwise be. It is possible that such programs contain provisions for persons with little or no prior economics exposure to be "brought up to speed" before being subjected to graduate level instruction. But we believe that it is highly unlikely that someone with no prior background in economics can be brought up to speed, be exposed to an appropriate quantity of graduate work at a sufficient level of rigor, and be able to complete the degree program in the 1 to 2 years that most institutions report as the normal time for completing the degree. It seems to us that either the depth of learning in master's programs is being sacrificed or that the content of the undergraduate degree in economics is lacking. In either case, we are concerned.

Admissions prerequisites seem to have deteriorated since the Bowen study in the early 1950s. Of the 84 departments for which data were obtained at that time, 46 percent required that the applicant have completed an undergraduate major in economics and only 7 percent had no prior substantive requirements in economics (Bowen, p. 80). Even so, Bowen found that one-third of all graduate students in economics majored in another field as undergraduates. There was even at that time a tendency to assume that students began their graduate careers with little knowledge of economics.

Course Requirements

The respondents to our questionnaire were asked to indicate those courses specifically required in their master's programs. The results that we gathered were surprisingly diverse and indicate that program uniformity, as reflected in required courses, is limited to four analytical areas. Microeconomic theory and macroeconomic theory are required in nearly all reporting programs—97 percent and 96 percent respectively. A course in statistics and/or econometrics is required in 82 percent of all responding institutions, while mathematical economics is required in 60 percent. Of equal interest, perhaps, is the fact that a few courses which some might have thought would be commonly required are not. For example, only 11 percent cited history of economic thought as a required course, and only 4 percent mentioned economic history.

Respondents listed a number of other courses as required in their specific programs, but none so frequently as to comprise a "common core" for master's programs across institutions. Among such courses were research methods, international economics, monetary economics, and forecasting. But beyond the two-course "common core" of microeconomic and macroeconomic theory and the frequent additional requirements of statistics/econometrics and mathematical economics, most master's programs appear to consist primarily of elective courses.

The course requirements of master's degree programs used to be more diverse, as is shown in Table 1. But the trend toward today's theory-plus-electives pattern was discerned by Bowen 35 years ago. After asking what courses *ought* to be required in a "common core," he reported (Bowen, p. 105): "Almost everyone agrees that there should be a common core, but economic theory is the only subject which all agree should be included."

Comprehensive Examinations

In our survey, a majority (62 percent) of institutions with master's programs in economics indicated that they had either a requirement of or an option for at least one comprehensive examination (field examination). In approximately half of these institutions the comprehensive examination was a requirement, while in most of the other cases the comprehensive examination was an option to writing a thesis, with one or the other being required. When we asked the respondents to indicate the areas of the comprehensive examinations, we discovered that 30 percent of those institutions with provisions for comprehensive examinations offered an examination in economic theory only. Further, 59 percent of the remaining institutions with provisions for comprehensive examinations offered such exams in economic theory *and* another area or areas. The requirements varied considerably, but the typical requirement was one of an examination in economic theory and an examination in one applied field or

Table 1

Selected course requirements in economics master's programs, 1985–86 vs. 1951–52

	Percent of reporting institutions in which course is formally required	
Course	*Thornton & Innes survey 1985–86 (n = 152)*	*Bowen survey 1951–52 (n = 59)*
Microeconomic theory	97%	63%[a]
Macroeconomic theory	96%	
Statistics and/or econometrics	82%	16%[b]
Mathematical economics	60%	1%[b]
History of economic thought	11%	25%
Economic history	4%	6%

Source: Bowen, p. 105.

[a] In the Bowen survey, economic theory was listed as a single course or field rather than broken into its micro and macro components.

[b] Bowen's survey listed statistics separately and pooled the econometrics and mathematical economics responses.

major area of specialization. A smaller number of programs indicated that exams in economic theory and a quantitative area (for example, econometrics, statistics, mathematical economics) were required.

Unfortunately, it is not possible to state definitively whether the comprehensive examination feature of master's programs is on the wane. While Bowen reported that "most departments provide comprehensive examinations" (pp. 207–8), he offered no quantitative information.

The Master's Thesis

The case of the master's thesis reveals the most striking evidence of change in the structure of master's programs. Bowen reported that of the 107 universities for which information was available, 86 percent required a master's thesis (or a master's essay) in 1951–52. In our own survey, only 24 percent of institutions with programs reported a thesis requirement. However, although only a minority of institutions require a thesis, it exists as an option in many other institutions. For example, our survey found that 30 percent of master's programs had a requirement of either a thesis or a comprehensive examination. Another 10 percent of institutions reported that there was

Table 2

Average time spent by students pursuing master's degree full time, 1985–86

Number of years	*Percent of responding institutions*[a]
1	15.7%
1.5	56.9%
2.0	25.5%
other	1.9%

[a] Information refers to the percent of institutions indicating the corresponding number of years a typical full-time student takes to complete the program.

an unrestricted option for a thesis: that is, an option which was not required as an alternative to a comprehensive exam.[3]

The foothold of the master's thesis was already beginning to weaken at the time of Bowen's survey. Despite noting a high degree of support and satisfaction among both graduate students and faculty with the purpose and functioning of the thesis, Bowen also noted what is probably the most important reason for its subsequent gradual demise—the very considerable burden of supervision which it imposes on faculty (especially in departments with large graduate enrollments).

Time Spent and Post-Master's Pursuits

It was not our purpose to create a profile of students pursuing the master's degree in economics. However, we did gather some information about the average time spent pursuing the degree and the subsequent career choices of master's students.

Approximately 60 percent of students enrolled in the typical master's degree program today are full-time students. As Table 2 indicates, most institutions reported that a typical full-time student can probably expect to spend between 1 1/2 and 2 years in pursuit of the degree, approximately the same average time that Bowen found that students spent between the bachelor's and master's degree (p. 6).

To determine the post-degree pursuits of students in master's programs, we asked the respondents to our questionnaire to indicate which of the four pursuits was the most typical "next step" taken by graduates immediately after completion of their master's program: business employment, government employment, Ph.D. program, or teaching employment. Table 3 contains the responses, but one caution is in order. The percentages in the table indicate the (unweighted) institutional response to the alternatives offered, and not the percentages of students entering the various pursuits.

[3] Although we did not specifically request the information, a small number of respondents indicated that a "master's essay" or "master's paper"—as opposed to a thesis—was either a program requirement or option. The distinctions offered between a master's thesis and a master's essay varied somewhat. Some respondents indicated that the thesis was "original research" and the essay was not, while others noted that the completed essay—unlike the thesis—was not deposited in the library.

Table 3
Most typical next step pursued by graduates of economics master's programs

Typical next step	*Percent of responding institutions*
Business employment	50%
Government employment	23%
Ph.D. program	25%
Teaching employment	2%

Note: Information refers to the percent of institutions indicating the most typical next step pursued by graduates of economics master's programs.

Nevertheless, the responses still clearly indicate that the master's in economics is generally a terminal degree, at least from the vantage point of the average institution offering the programs.

Nearly 75 percent of all institutions responded that most of their master's degree recipients accept employment with business or government rather than continue on toward a doctorate. Teaching employment, to the surprise of no one, was the least popular response. This finding, of course, reflects the fact that the Ph.D. is a nearly universal requirement in higher education today.

Bowen's survey of 35 years ago provided limited information on the employment patterns of a very small sample of 54 persons who received master's degrees over the decade preceding 1950. About 25 percent of this group were employed in higher education, while about 58 percent held jobs with noneducational public or private agencies (p. 10).

The Master's Degree as a "Consolation Degree"

When the returns from our survey were tabulated, we were surprised to find that 15 respondents indicated that they had no master's programs, contrary to the information provided in Peterson's *Graduate Programs*. The apparent discrepancy is explained by the fact that a number of institutions award master's degrees to individuals who have been enrolled in Ph.D. programs but who have failed to complete the requirements. For these schools it is probably fair to say that the master's degree is a "consolation degree." The responses of the remaining schools in our survey that offer both Ph.D. and master's degrees indicate that a majority also engage in this practice. Of the 83 respondents whose departments offer both the master's and Ph.D. degrees, 44 (53 percent) answered "yes" to the question: "In the event of a student failing to pass the requirements for a Ph.D. degree, is the master's degree in economics sometimes awarded as a 'consolation degree'?"

Many respondents added qualifying comments to their answers. These ranged from the unequivocal "frequently" to the ambiguous "if student meets requirements."

In fact, a small number of respondents who answered "no" to this question (or did not answer it all) offered similar comments. This suggests that the distinction between those who do offer a "consolation" master's degree and those who do not may be somewhat blurred. All told, however, our results indicate that about 60 percent of the Ph.D. granting institutions in our survey sometimes award such master's degrees.

It is unclear whether the practice of awarding "consolation" master's degrees is more common today than it was thirty years ago, since neither Bowen's study nor our own reports on the frequency with which such degrees have been granted.

Conclusions

The net effect of the above findings is somewhat disturbing. The picture of the master's degree contained in these findings is hardly in accord with Bowen's vision or with our own generalized ideal. Indeed, the necessary elemental aspects of such a master's program are less in evidence today than they were thirty-five years ago. Prerequisites for admission appear to have been lowered, the common body of knowledge required remains minimal, and the writing of a thesis to demonstrate professional competence is becoming rare. We can only conclude that the master's degree has been devalued, relative to the bachelor's and doctor's degrees. Our finding that the degree is often used as a "consolation degree" may simply be symptomatic of this devaluation, though it is likely that it is also a contributing factor. However, we suspect that Bowen's view of the master's degree, as one which should signify "a substantial accomplishment in economics and a level of professional competence as an economist," is a view which is still widely shared today. If this is indeed the case, then the erosion of standards reflected in our survey results must be halted and reversed.

Perhaps it is time for the economics profession to undertake another comprehensive study of its graduate programs—both master's and Ph.D. Such a study is warranted by its potential revelations in many other areas than those discussed here. In any case, the fact remains that it has been three decades since the last such comprehensive study was undertaken, and this reason alone is sufficiently compelling.

■ *We wish to thank Howard R. Bowen, W. Lee Hansen, John J. Siegfried, and the editors for helpful comments.*

References

Barr, S., and G. Young, "A Survey of Graduate Economics Programs in the United States," paper presented at the annual meetings of the Eastern Economics Association, 1984.

Bowen, H. R., "Graduate Education in Economics," *American Economic Review*, supplement, September 1953, *43:4*, part 2.

Graduate Programs in the Humanities and Social Services, 19th ed., Peterson's Annual Guides/Graduate Study, 1985.

Sterrett, J., and S. Barr, "Graduate Economics Programs in the United States," *Journal of Economic Education*, Fall 1983, 65–67.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 179–185

Puzzles
Penny Stocks, Discount Brokers, Better Bidding, and More

Barry Nalebuff

In presenting economic puzzles, I have three goals in mind: some puzzles are chosen to stimulate research; others offer examples that will help undergraduate and graduate teaching; all should provide quality distractions during seminars. This feature begins with reader's comments on the first "Puzzles." Next are several speed puzzles, with answers provided at the end. Here we have two Fischer Black problems in finance and a dot game from Richard Zeckhauser. Following are several longer puzzles, for which readers are invited—nay, challenged—to submit their own answers. This issue's longer problems give you a chance to think about strategies for staying ahead (in sailboat racing to R&D racing) and not losing your head at an auction. The responses will be discussed in a future issue.

Please send your answers and favorite puzzles to: Barry Nalebuff, "Puzzles," c/o *Journal of Economic Perspectives*, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton NJ 08544. Good luck.

Mail

It was inevitable that my first letter would point out a mistake. The proposed answer to Puzzle 2 ("In Fact, it's a Gas") in the first issue, concluded that "self-service gasoline is outlawed in New Jersey (and Alaska)." Professor Gunnar Knapp at the University of Alaska, Anchorage writes that New Jersey has that distinction all to

■ *Barry Nalebuff is Assistant Professor of Economics, Princeton University, Princeton, New Jersey.*

itself. This raises a new puzzle: why is self-service gas outlawed anywhere? Are there any special interest groups that benefit from this legislation? What is special about New Jersey? On the positive side, Professor J. Wilson Mixon from Berry College, Georgia, writes that his econometric tests (Mixon and Uri, 1987) support the hypothesis that the price premium for unleaded gasoline is due to differential price elasticities between consumers of leaded and unleaded gas.

Puzzle 1: A Penny Saved is a Penny Burned?

Imagine that the price of copper rises to the point where the copper value of a 1 cent coin is worth more than a penny. As a result, pennies disappear from circulation and no one misses them.

Your firm uses copper in its production process, and you can melt pennies down and retrieve their copper content at zero cost. At present, you have a six-month stockpile of copper reserves and you have managed to collect one million pennies. Should you melt the pennies down? If so, why, and if not, why not?

Puzzle 2: You Get What You Pay For

An investment adviser offers to manage your portfolio for a fee. In return for her investment expertise, she asks for 10 percent of the excess return (that is, the return above a safe asset, perhaps the short term treasury bill rate). If by chance she underperforms the safe asset, then she will also return to you 10 percent of the amount that she lost (relative to the safe asset).

As an alternative fee structure, the investment counselor offers to accept a flat percentage of the overall portfolio value. What yearly percentage of the portfolio would be equivalent, in your eyes, to her proposed fee of 10 percent of the excess return?

Puzzle 3: This Dot is For You

Any game in which players move in sequence and which always ends after a finite number of moves has a "best strategy." In theory, this best way of playing could be discovered by examining every possible sequence of moves. This approach is relatively easy for tic-tac-toe and impossible (at present) for chess. In the game below, the best strategy is unknown. Yet, even without knowing it, the fact that it exists is enough to show that it must lead to a win for the first player.

Richard Zeckhauser describes the following two player dot-game. The game starts with dots arranged in any rectangular shape. The object of the game is to force your opponent to take the last dot. For example, look at the 7 × 4 array below.

. . . .

. . . .

. . . .

. . . .

Each turn a player removes a dot and with it *all* remaining dots to the northeast. In our example, if the first player chooses the fourth dot in the second row, this leaves the second player opponent with

. . .

. . .

. . . .

. . . .

Each play, at least one dot must be removed. The person who is forced to take the last dot loses.

For any shaped rectangle with more than one dot, the first player has a winning strategy. Although this strategy is known for special cases (including the 7 × 4) a general winning strategy remains to be discovered. How can we know who has the winning strategy without knowing what it is?

Answers to Puzzles 1, 2, and 3 appear following Puzzle 5.

Puzzle 4: Keeping One Step Ahead

After the first four races in the 1983 America's Cup finals, Dennis Conner's *Liberty* led 3-1 in a best out of seven series. On the morning of the fifth race, "cases of champagne had been delivered to *Liberty's* dock. And on their spectator yacht, the wives of the crew were wearing red-white-and-blue shorts, in anticipation of having their picture taken after their husbands had prolonged the United States' winning streak to 132 years."[1] It was not to be.

At the start, *Liberty* got off to a 37-second lead when *Australia II* jumped the gun and had to recross the starting line. The Australian skipper John Bertrand tried to catch up by sailing far to the left of the course in the hopes of catching a wind shift. Dennis Conner choose to keep *Liberty* on the right hand side of the course. Bertrand's gamble paid off. The wind shifted five degrees in *Australia II*'s favor and they won the race by one minute and forty-seven seconds. Two races later, *Australia II* won the series.

---

[1] The *New York Times*, September 22, 1983, p. B19.

Conner was criticized for his strategic failure in the race. In sailboat racing, the leading sailboat usually follows the route of the trailing boat, even when the follower is clearly pursuing a poor strategy. Why? Because in sailboat racing, overall speed doesn't count. Only winning matters. If you have the lead, the surest way to preserve the lead is to play monkey see, monkey do.

While you might never get the chance to skipper an America's Cup race, here you have a chance to apply the same principles in a research and development race.

Going into the home stretch of an R&D project, you are six months ahead of the competition. To bring the project to a successful completion requires finishing the development stage. There are two strategies you can pursue, RISKY and SAFE. SAFE takes two years but is guaranteed to work. RISKY takes only one year, but there is a fifty percent chance that it will get you nowhere, in which case you will have to return to the SAFE strategy and take an additional two years.

Six months from now, your competitor will be faced with a similar problem of how to bring their development stage to completion. They too have a SAFE and a RISKY strategy available. Their SAFE strategy takes them two years to reach completion. Their RISKY strategy takes only one year, but has only a fifty percent chance of success. The competitor's chance of success is *independent* of whether your RISKY strategy works. When their RISKY strategy fails, they must turn to their SAFE option and wait an additional two years.

Only the first to completion is awarded the patent. There is no time discounting. Due to limited resources, it is not possible to pursue both strategies simultaneously. Which strategy should you pursue to maximize your chance of winning? Do you want to keep your move hidden? What do you think your competitor will do? How likely are you to win?

This problem differs from the sailboat racing example in one important way. The two sailboats can observe each other. It is relatively straightforward for Dennis Conner to follow John Bertrand. Staying ahead is more complicated when the moves are unobserved; prediction rather than observation is needed. It's your move.

Puzzle 5: Be a Better Bidder

One of the more interesting developments in auction theory is the revenue equivalence theorem.[2] This result demonstrates that a wide range of auction formats all provide the seller with the same expected revenue. Equivalence arises when there is a single indivisible object is for sale, the bidders do not collude, each bidder knows his own valuation of the object but not the value to others, and these private valuations are independently and identically distributed.

---

[2] Several examples of equivalence were first noticed by Vickrey (1962); the general equivalence theorems are due to Myerson (1981) and Riley and Samuelson (1981).

One type of problem where the revenue equivalence theorem does not apply is when the bidders have a common valuation for the sale object. For example, the value of an offshore oil lease may be the same for all bidders, but they each could have different estimates of this common value. This auction problem is characterized by a "winner's curse" and it is sometimes thought that this curse is what destroys revenue equivalence.[3]

In the problem below, Jeremy Bulow describes an auction where the bidders' valuations are still affiliated (as in the oil example) but where there is no winner's curse. Although each bidder has a private valuation and these values are identically distributed, different auctions result in different levels of expected payments to the seller. Hence it is not just the winner's curse which destroys revenue equivalence.[4]

There are two people bidding. Person 1 values the object at A + B. Person 2 values the object at B + C. The values A, B, and C are all independently and uniformly distributed on $[0, 1]$. Finally, each bidder only knows his total valuation; person 1 only knows the sum of A + B while person 2 only knows the sum of B + C.

It is worth emphasizing that there is no "winner's curse" in this problem. In this auction, both bidders know their own valuation exactly and their estimate is unaffected by the other's valuation. Now consider three types of auctions.

First, a sealed bid auction: the highest bidder wins the object and pays his bid.

Second, a second price sealed bid auction: the highest bidder wins the object but pays only the second highest bid.

Third, an auction where both bidders pays their bid, regardless of who wins, and the highest bidder is awarded the object.

What are the seller's expected revenues from each of these three auction formats? Why do they differ? Can you predict in advance which will work best? How much of the buyer's surplus can the seller get?

Compare these outcomes to a common value auction problem. Let the value to both bidders be A + B + C, where bidder 1 observes only A + B and bidder 2 observes only B + C. Now which auction works best for the seller?

I will report the answers to Puzzles 4 and 5 in a subsequent issue.

Answers to Speed Problems

Answer to Puzzle 1

Melting down copper pennies is never a good idea as long as you have some other source of copper already available. For that matter, it's not important, except for transportation costs, that you have the copper. So long as anyone has copper, if transportation costs are zero, it pays to hoard pennies and not melt them down. Why? If the price of copper stays high or rises further, it won't make any difference whether you melt the pennies or not. If the price of copper falls, however, the price of copper stockpile will decline but the price of the pennies will not fall to less than one cent each. Pennies can be turned into copper, but you can't mint the copper back into pennies. Because the value of pennies has a lower bound of one cent each, they are worth more than their copper value alone. Once you have melted the pennies, this surplus is lost. If the choice is between melting already-purchased pennies or buying more copper, and copper costs more than pennies, then pennies might be worth melting for immediate use. But as long as the copper stockpile last, it is always better to save or sell the pennies.

---

[3] The winner's curse arises when the high bidder discovers that everyone else had a lower estimation of the sale good's value. This then leads him to lower his expectation of the good's value. Of course, each bidder places his bid accounting for the curse, as his bid counts only if it is the highest. For a more detailed discussion, Richard Thaler describes the anomaly of the winner's curse in this issue.

[4] Milgrom and Weber (1981) provide the now standard treatment of symmetric buyers with affiliated values, and this problem is meant to help illustrate some of their results.

This problem is just a special case of Bob Merton's (1973) result that it never pays to exercise early a stock option that doesn't pay dividends. Pennies are like copper with a free option.

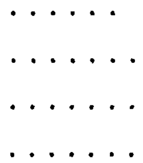
Answer to Puzzle 2

The broker who is willing to accept 10 percent of the *excess* return (positive or negative) provides her services for *free* in a world where one can borrow and lend at the risk-free rate. One way of understanding this result is to recognize that 10 percent of the excess return on the entire portfolio is equivalent to all of the excess return on 10 percent of the portfolio. So from the investor's point of view, the offer of the broker to take 10 percent of excess return on a $100 investment is equivalent to putting $90 into the market and $10 into risk-free Treasury bonds. If you would have preferred putting all $100 into the market, then borrow $11 and give the broker $111. After she takes the 10 percent of the excess return, she will give you the equivalent of the risk-free rate on $11 (which you use to pay back your loan) and the market return on the other $100. Of course, in this world, the broker doesn't need your money. She could just as easily borrow money from the bank and offer to keep 100 percent of the excess return, leaving them with just the risk-free rate that they require.

The broker's services are costly to the extent you cannot borrow at the risk-free rate. But as long as you are not up against that constraint, her services are free under the first scheme.

Answer to Puzzle 3

If the second player in the game has a winning strategy, that means that for *any* opening move of the first player, the second has a response that puts him in a winning position. Imagine that the first player takes just the upper right hand dot.



No matter how the second player responds, the board will be left in a position that the first player could have chosen as a first move. If this is truly a winning position, the first player should have and could have opened the game this way. There is nothing the second player can do to the first that the first player cannot do unto him beforehand.

■ *The first two speed puzzles are adapted from a course Fischer Black taught at MIT (15.423). Thanks to Pete Kyle for reminding me of two favorites. Problems 3 and 4 are adapted from my forthcoming book with Avinash Dixit,* Strategic Thinking and Action. *Jeremy Bulow deserves the credit for Problem 5.*

References

Merton, Robert, "The Theory of Rational Option Pricing," *Bell Journal of Economics*, Spring 1973, *41*, 141–183.

Milgrom, Paul, and Robert Weber, "Theory of Auctions," *Econometrica*, September 1982, *50*, 1089–1121.

Mixon, J. Wilson, and Noel Uri, "Price Discrimination and Free-entry Market: the Case of Gasoline," *Applied Energy*, 1987, *26*, 1–7.

Myerson, Roger, "Optimal Auction Design," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

Riley, John, and William Samuelson, "Optimal Auctions," *American Economic Review*, June 1981, *71*, 381–92.

Vickrey, William, "Auctions and Bidding Games," *Recent Advances in Game Theory* (Conference Proceedings), Princeton University, 1962, 15–27.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 187–189

Recommendations for Further Reading

Bernard Saffran

This section lists readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, the articles chosen are expository or integrative and do not focus on original research. As a rule of thumb, articles mentioned here appear in journals not indexed by the *Journal of Economic Literature* (which focuses on the research journals) and are similar to many of the articles published in this journal. They may include survey articles, discussion of related subjects in which economists might have an interest (for example, the use of game theory in evolutionary biology) or analyses of economics from other perspectives.

The intention is to publish a selective list of ten to fifteen articles per issue, with the selections depending heavily on input from readers. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Bernard Saffran, c/o *Journal of Economic Perspectives*, Department of Economics, Swarthmore College, Swarthmore, Pennsylvania 19081.

Paper Potpourri

These papers range from equal pay to productivity, from income distribution to income maintenance, and from economical writers to Keynesian economic thought.

Bergmann, Barbara R., "Pay Equity—Surprising Answers to Hard Questions," *Challenge*, May-June 1987, 45–51. A human capital theory argument for using comparable worth as a criterion for setting pay.

Griliches, Zvi, "R&D and Productivity: Measurement Issues and Econometric Results," *Science*, July 3, 1987, 31–35. "A recent study finds a significant contribution

■ *Bernard Saffran is Professor of Economics, Swarthmore College, Swarthmore, Pennsylvania.*

of R&D to productivity growth in the largest U.S. manufacturing corporations, with no evidence of a major decline in it, and a larger role for basic research and a smaller one for federally financed R&D expenditures than is implied by their relative importance in total R&D expenditures."

High, Jack C., "The Costs of Economical Writing" and McCloskey, Donald N., "Reply to Jack High," *Economic Inquiry*, July 1987, 543–548. A comment and reply on the private costs of good writing. The discussion is based on McCloskey's earlier article "Economical Writing," which was mentioned in the Summer 1987 of this journal.

Kindleberger, Charles P., "Fun Reading for Bankers," *Forbes*, July 13, 1987, 292+. A witty, annotated bibliography of fifty books that includes a personal evaluation of many of history's leading bankers and monetary analysts. This bibliography is fun to read even if you do not get around to reading all of the books on the list.

Levy, Frank, "Changes in the Distribution of American Family Incomes, 1947 to 1984," *Science*, May 22, 1987, 923–927. A discussion of two major controversies: the reasons for the unchanged income distribution since 1947 and whether there has been a change in living standards since 1974. "These controversies can be partially resolved by understanding the sharp slowdown in the growth of workers' wages which occurred after 1973 and the demographic trends which kept per capita living standards rising despite stagnant wages, including more working women and low birthrates."

Munnell, Alicia H., "Lessons from the Income Maintenance Experiments: An Overview," *New England Economic Review*, Federal Reserve Bank of Boston, May/June 1987, 32–45. In the belief that "the major lessons of the experiments are neither widely known nor well understood" and to "summarize the findings and to derive the methodological and policy lessons" the Federal Reserve Bank of Boston and The Brookings Institution sponsored this conference in the fall of 1986 at which economists and other social scientists evaluated the income maintenance experiments. This article is a summary. The Proceedings are available without charge (as conference series No. 30) from Research Library-D, Federal Reserve Bank of Boston, Boston, Massachusetts 02106.

Tobin, James, "The Future of Keynesian Economics," *Eastern Economic Journal*, October-December 1986, 347–356. After discussing the recent attacks on it, Tobin concludes, "Yes, Keynesian economics has a future because it is essential to the explanation and understanding of a host of observations and experiences, past and present, that alternative macroeconomic approaches do not illuminate." Of particular interest are the Appendices which consist of two amusing poems "Adrift in the Islands or Unreal Business Cycle Theory" and "The Shocking Game the Gods Play or Real Business Cycle Theory".

History of the Profession

These articles describe how Europe affected American economists and how economists helped to effect social reform.

Cookingham, Mary E., "Social Economists and Reform: Berkeley, 1906-1961," *History of Political Economy*, Spring 1987, 47-65. While the Berkeley economists' impact on the social reform movement was not as great as that of the economists at Wisconsin, this article should still be of wide interest because of its discussion of the work of the three women tenured faculty members, Jessica Peixotto, Lucy Stebbins and Emily Huntington.

Craver, Earlene, and Axel Leijonhufvud, "Economics in America: the Continental Influence," and Solow, Robert M., "What Do We Know That Francis Amasa Walker Didn't?" *History of Political Economy*, Summer 1987, 173-190. Papers given at the 1985 A.E.A. meetings for a session titled "American Economics Officially Turns One Hundred."

Green Grass from Neighboring Fields

At worst, neighboring academic fields serve as a prod to inspiration and a source of metaphor. But insights from other fields like sociobiology or weather forecasting can also help to develop purely economic insights.

Kitcher, Philip, "Précis of *Vaulting Ambition: Sociobiology and the Quest for Human Nature*," *Behavioral and Brain Sciences*, March 1987, 61-100. This is a useful introduction to the debate on the merits of sociobiology. Kitcher is a philosopher at the University of California at San Diego. The first ten pages of the article summarize his new book (published by MIT Press in 1985), in which he "seeks to identify what is good in sociobiology, to expose the errors of premature speculations about human nature, and to prepare the way for serious study of the evolution of human social behavior." This is followed by comments from scholars in many fields, a rejoinder by Kitcher and a bibliography.

Tribbia, Joseph J., and Richard A. Anthes, "Scientific Basis of Modern Weather Prediction," *Science*, July 31, 1987, 493-499. A review of some of the methodological issues in weather forecasting. Since both meteorology and economics are still largely nonexperimental sciences, many of the methodological difficulties in weather forecasting may apply to economic forecasting as well. The authors are with the National Center for Atmospheric Research.

Personal Touches

There have been a number of articles that are autobiographical and also contain avuncular advice.

Bergson, Abram, "Recollections and Reflections of a Comparativist," *The American Economist*, Spring 1987, 3-8.

"A Conversation with Martin Bronfenbrenner," *Eastern Economic Journal*, January-March 1987, 1-6.

"Interview with Paul M. Sweezy," *Monthly Review*, April 1987, 1-28.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 191–202

Anomalies
The Winner's Curse

Richard H. Thaler

Economics is distinguished from other social sciences by the belief that most (all?) behavior can be explained by assuming that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear. An empirical result qualifies as an anomaly if it is difficult to "rationalize," or if implausible assumptions are necessary to explain it within the paradigm. This column will present a series of such anomalies. Of course, "difficult" and "implausible" are judgments, and others might disagree with my assessment. Therefore, I invite readers to submit *brief* explanations (within the paradigm or otherwise) for any of the anomalies I report. To be considered for publication, however, proposed explanations must be falsifiable, at least in principle. Future topics for this column will come from as many fields of empirical economics as possible. Readers are invited to suggest topics by sending a note with some references to (or better yet copies of) the relevant research. My address is: Richard Thaler, c/o *Journal of Economic Perspectives*, Johnson Graduate School of Management, Malott Hall, Cornell University, Ithaca, NY 14853.

Introduction

Next time that you find yourself a little short of cash for lunch, try the following experiment in your class. Take a jar and fill it with coins, noting the total value of the

■ *Richard H. Thaler is Professor of Economics, Johnson Graduate School of Management, Cornell University, Ithaca, New York.*

coins. Now auction off the jar to your class (offering to pay the winning bidder in bills to control for penny aversion). Chances are very high that the following results will be obtained: (1) the average bid will be significantly less than the value of the coins (bidders are risk averse); (2) the winning bid will exceed the value of the jar. Therefore, you will have money for lunch, and your students will have learned first-hand about the "winner's curse."

The winner's curse is a concept that was first discussed in the literature by three Atlantic Richfield engineers, Capen, Clapp, and Campbell (1971). The idea is simple. Suppose many oil companies are interested in purchasing the drilling rights to a particular parcel of land. Let's assume that the rights are worth the same amount to all bidders, that is, the auction is what is called a *common value* auction. Further, suppose that each bidding firm obtains an estimate of the value of the rights from its experts. Assume that the estimates are unbiased, so the mean of the estimates is equal to the common value of the tract. What is likely to happen in the auction? Given the difficulty of estimating the amount of oil in a given location, the estimates of the experts will vary substantially, some far too high and some too low. Even if companies bid somewhat less than the estimate their expert provided, the firms whose experts provided high estimates will tend to bid more than the firms whose experts guessed lower. Indeed, it may occur that the firm that wins the auction will be the one whose experts provided the highest estimates. If this happens, the winner of the auction is likely to be a loser. The winner can be said to be "cursed" in one of two ways: (1) the winning bid exceeds the value of the tract, so the firm loses money; or (2) the value of the tract is less than the expert's estimate so the winning firm is disappointed. Call these winner's curse versions 1 and 2 respectively. Notice that the milder version 2 can apply even if the winning bidder makes a profit, as long as the profit is less than expected at the time the bid was made. In either version the winner is unhappy about the outcome, so both definitions seem appropriate.

The winner's curse cannot occur if all the bidders are rational (see Cox and Isaac, 1984), so evidence of a winner's curse in market settings would constitute an anomaly. However, acting rationally in a common value auction can be difficult. Rational bidding requires first distinguishing between the expected value of the object for sale, conditioned only on the prior information available, and the expected value conditioned on winning the auction. Even if a bidder grasps this basic concept, version 2 of the winner's curse can occur if the bidder underestimates the magnitude of the adjustment necessary to compensate for the presence of other bidders.

In a first price auction there are two factors to consider, and they work in opposite directions. An increase in the number of other bidders implies that to win the auction you must bid more aggressively, but their presence also increases the chance that if you win, you will have overestimated the value of the object for sale—suggesting that you should bid less aggressively.[1] Solving for the optimal bid is not trivial. Thus, it is an empirical question whether bidders in various contexts get it right or are cursed. I will present some evidence, both from experimental and field studies, suggesting that the winner's curse may be a common phenomenon.

---

[1] As Capen, Clapp and Campbell state (p. 645): "If one wins a tract against two or three others he may feel fine about his good fortune. But how should he feel if he won against 50 others?" "Ill."

Experimental Evidence

The jar of coins example cited above has, in fact, been conducted under experimental conditions by Max Bazerman and William Samuelson (1983). Their subjects were M.B.A. students taking microeconomics classes at Boston University. The objects auctioned off were jars of coins or other objects such as paper clips valued at four cents each. Unknown to the subjects, each jar had a value of $8. Subjects submitted sealed bids and were told that the highest bidder would receive the defined value of the object less his or her bid. A total of 48 auctions were conducted, 4 in each of 12 classes. No feedback was provided until the entire experiment was completed. Subjects were also asked to estimate the value of each jar (point estimates and 90 percent confidence limits), and a $2 prize was offered for the best guess in each class.

The estimates of the actual values turned out to be biased downward. The mean estimate of the value of the jars was $5.13, well below the true value of $8.00. This bias, plus risk aversion, would tend to work against observing a winner's curse. Nevertheless, the mean winning bid was $10.01, producing an average loss to the winning bidder of $2.01. Clearly these experiments do not require large NSF grants!

Samuelson and Bazerman (1985) have run another series of experiments about the winner's curse in a different context. Try this problem yourself before continuing.

> In the following exercise, you will represent Company A (the acquirer) which is currently considering acquiring Company T (the target) by means of a tender offer. You plan to tender in cash for 100% of Company T's shares but are unsure how high a price to offer. The main complication is this: the value of the company depends directly on the outcome of a major oil exploration project it is currently undertaking.
>
> The very viability of Company T depends on the exploration outcome. In the worst case (if the exploration fails completely), the company under current management will be worth nothing—$0/share. In the best case (a complete success), the value under current management could be as high as $100/share. Given the range of exploration outcomes, all share values between $0 and $100 per share are considered equally likely. By all estimates the company will be worth considerably more in the hands of Company A than under current management. In fact, whatever the value under current management, the company will be worth 50 percent more under the management of Company A than under Company T.
>
> The board of directors of Company A has asked you to determine the price they should offer for Company T's shares. This offer must be made now, before the outcome of the drilling project is known.

> Thus, you (Company A) will not know the results of the exploration project when submitting your offer, but Company T will know the results when deciding whether or not to accept your offer. In addition, Company T is expected to accept any offer by Company A that is greater than or equal to the (per share) value of the company under its own management.
>
> As the representative of Company A, you are deliberating over price offers in the range $0/share to $150/share. What offer per share would you tender? (pp. 131–33).

The typical subject thinks about this problem roughly as follows: The firm has an expected value of $50 to Company T, which makes it worth $75 to Company A. Therefore if I suggest a bid somewhere between $50 and $75, Company A should make some money. This analysis fails to take into consideration the asymmetric information that is built into the problem. A correct analysis must calculate the expected value of the firm conditioned on the bid being accepted. If a bid B is accepted, then the company must be worth no more than B under current management for an average of $B/2$. Under new management, the average is 150 percent of this, or $3B/4$, which is still less than B, so it is best not to bid at all. Thus, this problem produces an extreme form of the winner's curse in which any positive bid yields an expected loss to the bidder.

This experiment was run in two conditions, one with monetary incentives and one without. The results, as shown in Table 1, are quite similar for the two conditions, with the bids in the condition with monetary incentives somewhat lower. In both conditions over 90 percent of the subjects make positive bids, and a majority are in the range between $50 and $75.

Economists often respond to examples like this by hypothesizing that although people can be fooled once or twice by such a problem, they will figure out the trap with experience. Sheryl Weiner, Max Bazerman, and John Carroll (1987) have investigated this hypothesis by giving the "buy-a-firm" problem to 69 Northwestern M.B.A. students via a microcomputer. All subjects repeated the experiment 20 times with financial incentives and feedback after each trial. The feedback included the "true" value of the company, whether their bid was accepted, and how much money they made or lost. Of the 69 subjects, 5 learned to bid one dollar or less by the end of the experiment. For these 5 subjects, the average trial in which they began to bid $1.00 or less was trial 8. There was no sign of any learning among the others; in fact the average bid drifted up over the last few trials. It may be possible to learn to avoid the winner's curse in this problem, but the learning is neither easy nor fast.

Another series of experiments has been conducted by John Kagel and his colleagues at the University of Houston. Many of the experiments have the following structure. An object is to be sold using a sealed bid auction. The value of the object, X^*, is drawn from a uniform distribution over the interval X_L to X_H. Each bidder is then given a single-valued draw, X_i, from the uniform distribution $X^* \pm \varepsilon$, with ε varying from trial to trial. Subjects thus know X_L, X_H, and ε, and their own draw, X_i, which corresponds to the expert estimate in the oil bidding example. An auction is

Table 1
Distribution of bids in Samuelson and Bazerman experiment

Bids	*No Incentives* (*N* = *123*)	*Monetary incentives* (*N* = *66*)
0	9%	8%
1–49	16	29
50–59	37	26
60–69	15	13
70–79	22	20
80 +	1	4

Source: Samuelson and Bazerman, 1985.

then conducted, the bids are posted, and the winner has a profit or loss credited to his or her account. (Bidders were given some capital to start with, usually about $10. Once their account reached zero they were no longer permitted to bid.) The experimental manipulations included varying ε, N (the number of bidders), and the type of auction (first price, second price, and low price[2]). Typically subjects first participated in small groups of 3–5 bidders and then in "large" groups of 6 or 7. A nice feature of all the experiments is that for each trial the authors compute the outcome that would be predicted by a bidding model using a noncooperative Nash equilibrium concept and risk neutral bidders. They call this the risk neutral Nash equilibrium or RNNE model.[3]

In Kagel and Levin (1986), which used first price auctions, the results varied with group size. In the small groups, there were typically profits which were, on average, 65.1 percent of the RNNE profits. However, in the large groups, losses of $.88 per auction period were observed, in contrast to the $4.68 profit predicted by the RNNE. The winner's curse emerged in the larger groups because the subjects bid more aggressively in the larger groups, while the RNNE bid function requires bidding more conservatively as N increases. These results have been replicated in Kagel, Levin and Harstad (1987) using a second price auction method. Again there were profits in the small group experiments, here 52.8 percent of RNNE profits, and losses of $2.15 per period in the large groups, compared to the RNNE prediction of $3.95 in profits.

Finally, Dyer, Kagel and Levin (1987) report on a series of low price auctions. In these auctions there were losses in both the small groups and the larger groups. However, the most interesting and innovative feature of the paper is that they included an experiment with a group of construction firm managers. A common criticism of experimental economics, especially if the results of the experiment do not accord with economic theory, is that the subjects are "only college students working on toy problems, and in the real world, experts wouldn't make these silly mistakes." So how did the construction managers do? While the experimenters worried that the experts might take them to the cleaners, in fact the experts did no better or worse than the students. This is surprising, given that construction firms participate in low bid auctions all the time, and would soon go bankrupt if they fell prey to the winner's curse. Dyer, Kagel and Levin believe the result occurs because the managers have learned situation specific rules of thumb rather than the relevant theory:

---

[2] In a first price auction, the item is awarded to the highest bidder who pays whatever he bid. In a second price auction, the item is given to the highest bidder, but he pays whatever was the second highest bid. In a low price auction, such as construction contract bidding, the winner is the one who submits the lowest bid. Low price auctions can use either first or second price rules.

[3] That is, each risk neutral bidder has a strategy that he would not want to change even if he knew all the other bidders' strategies.

> We believe that in the field the executives have learned a set of situation specific rules of thumb which permit them to avoid the winner's curse in the field but could not be applied in the laboratory.... However, these rules of thumb do not translate into a structurally similar, but different environment which lacks a familiar references. When placed in a new environment which does not contain the usual stimuli, the learning process must take place anew, since without theory absorption there is nothing to be carried over from previous experience (pp. 23-24).

Field Data

The laboratory evidence demonstrates that avoiding the winner's curse is not easy. Even experienced subjects who are given significant learning opportunities fail to solve the buy-a-firm problem and fail to understand the need to become more conservative when the number of bidders increases. Do bidders in large stakes auctions in the "real world" make the same mistakes? There are numerous studies that claim to have found evidence of the winner's curse in market contexts. For example, in the field of book publishing, Dessauer (1981, p. 33) reports that: "The problem is, simply, that most of the auctioned books are not earning their advances. In fact, very often such books have turned out to be dismal failures whose value was more perceived than real."[4] Cassing and Douglas (1980) looked at the market for free agents in baseball and concluded that free agents were overpaid. The owners of major league baseball teams seem to have come to the same conclusion, and appear to have responded with the effective tactic of collusion.[5] The fact that during the 1986-87 off-season, no team, including George Steinbrenner's pitching-weak Yankees, even bid for Jack Morris, perhaps the American League's best pitcher in the 1980s, is reasonably strong evidence of collusion. Here I will review the evidence in two other contexts: offshore oil and gas leases, and corporate takeovers.

---

[4] Of course, the quoted statement could be true but not be evidence of a winner's curse if the sales distribution is sufficiently skewed.

[5] It has also been suggested that the construction managers in Dyer, *et al.*'s experiments would have done better in an experiment emphasizing cartel skills rather than optimal bidding strategy.

It is appropriate to start with the evidence on bidding for oil and gas drilling rights, since that is the domain which prompted the marvelous article by Capen, Clapp and Campbell that first mentioned the concept of a winner's curse. They began their discussion by noting (p. 641):

> In recent years, several major companies have taken a rather careful look at their record and those of the industry in areas where sealed competitive bidding is the method of acquiring leases. The most notable of these areas, and perhaps the most interesting, is the Gulf of Mexico. Most analysts turn up with the rather shocking result that, while there seems to be a lot of oil and gas in the region, the industry is not making as much return on its investments as it intended. In fact, if one ignores the era before 1950, when land was a good deal cheaper, he finds that the Gulf has paid off at something less than the local credit union.

The authors cite several studies to document their claims, and report some interesting data of their own regarding the dispersion of bids. They report that the ratio between the highest and lowest bids by what they call "serious competitors" is commonly as high as 5 to 10 and can be as high as 100. While this result might be explained by some firms submitting low bids in the hope that there will be no other bidders (as was true for 15 tracts in the sample they analyzed), the authors report some other interesting data. In the 1969 Alaska North Slope sale the sum of the winning bids was $900 million, while the sum of the second highest bids was only $370 million. The winning bid exceeded the second bid by a factor of 4 or more in 26 percent of the tracts, and by a factor of at least 2 for 77 percent of the tracts. While these figures do not actually prove that anyone was behaving irrationally, they certainly seem consistent with a winner's curse scenario.

The Capen, Clapp and Campbell article was published in 1971, before all the information was in on the Gulf of Mexico leases they discuss. However, Walter Mead, Asbjorn Moseidjord, and Philip Sorensen (1983) have examined how those leases turned out. They calculate before-tax rates of return on 1,223 leases issued in the Gulf of Mexico between 1954 and 1969, the period directly preceding the publication of Capen, Clapp and Campbell. They report (p. 42):

> [F]or all 1223 leases, firms suffered an average present value loss of $192,128 per lease using a 12.5% discount rate.[6]... 62% of all leases in our data base were dry. Consequently, the lessees had no revenues whatsoever to offset their bonus and rent payments, or their exploration costs. Another 16% of the leases were unprofitable (on an after-tax basis) although some production occurred. Only 22% of the leases were profitable, and these leases earned only 18.74% in aggregate on the after-tax basis.

These results seem consistent at least with version 2 of the winner's curse; that is, they are surely lower than the bidders anticipated when they bid on the property. In addition, these returns are helped by the fact that nominal crude oil prices increased from $3 to $35 a barrel between 1970 and 1981, something that could not have been anticipated when the leases were purchased. As to why the returns were so low, the authors venture this (p. 45): "The low and negative rates of return for the initial five lease sales (from 10/13/54 through 8/11/59) appear to reflect excessive enthusiasm for the amount of oil likely to be found."

---

[6] They use nominal values for costs and selling prices; thus this rate seems reasonable.

Another analysis of the same lease sales was conducted by Hendricks, Porter, and Boudreau (1987). They use a 5 percent real discount rate and real price sequences that did not assume that the oil companies could anticipate the OPEC price shocks. They also make several other assumptions that differ from those used by Mead, Moseidjord and Sorenson. Their results, in contrast, indicate that firms would have made profits even if oil prices had remained constant in real terms. Nevertheless, their data does provide some support for the winner's curse. For the 18 individual firms or consortiums of firms that made a significant number of bids (the average number of bids was 225), Hendricks, *et al.* calculated the profit each firm would have made ex post if it had multiplied all of its bids by a constant, θ, assuming all other firms had kept their bids the same. They then determined θ^*, the value of θ that would have maximized profits. If all firms chose their bids according to risk neutral Nash equilibrium behavior, then θ^* would be equal to one. However, for 12 of the 18 firms θ^* was less than one, with the median value being .68. For Texaco, which seems to have been particularly cursed, θ^* was .15, indicating that they should have reduced their bids by a factor of nearly 7! For many of the firms, the difference between the actual profits earned and those that would have been earned with optimal bidding amounted to hundreds of millions of dollars. The authors conclude (p. 529): "This result suggests that some firms may have systematically overvalued the tracts and/or failed to fully anticipate the impact of the 'winner's curse'."

Richard Roll (1986) applies the concept of the winner's curse to the puzzling phenomenon of corporate takeovers. The puzzle is to explain why firms are willing to pay substantial premiums above the market price to acquire another firm. The empirical evidence suggests that while stockholders of target firms make significant profits when their firms are purchased, there is little or no gain to the buyer. Why then do takeovers occur? Roll offers what he calls the hubris hypothesis as one plausible answer. According to this view, bidding firms, typically flush with cash[7] identify potential target firms, estimate the value of the target, then bid for the target if and only if the estimated value exceeds the market value. Since Roll takes the efficient market hypothesis seriously, he believes that (in the absence of synergy or

[7] Asquith (1983) reports that successful bidders earned 14.3 percent above the market over the 460-day period ending 20 days before the merger. In light of this fact, I think that the hubris hypothesis may be partially a "hot hand" phenomenon. Most basketball players and fans believe that there is strong positive serial correlation in the shooting of basketball players; that is, the probability of making a shot increases after making the previous shot and vice versa. In contrast to this perception, psychologists Gilovich, Vallone, and Tversky (1985) have found no serial correlation using actual NBA data. So according to the hot-hand-hubris hypothesis, firms that have been doing well lately, perhaps because of a run of good luck, mistakenly think that they are "hot" (i.e., good managers) and will be able to perform miracles with any firm they purchase.

insider information) the belief by acquirers that they can estimate true value better than the market is likely to be mistaken. As Roll points out (p. 201):

> Most other explanations of the takeover phenomenon rely on strong-form market inefficiency of at least a temporary duration. Either financial markets are ignorant of the relevant information possessed by bidding firms, or product markets are inefficiently organized so that potential synergies, monopolies, or tax savings are being ineffectively exploited (at least temporarily), or labor markets are inefficient because gains could be obtained by replacement of inferior managers.

To test the hubris hypothesis Roll reviews the data on stock prices of bidders and targets around the announcement date. The hubris hypothesis predicts that the combined value of the bidder and the target should fall slightly, representing transactions costs; the value of the target firm should increase; and the value of the bidder should fall.[8] He interprets the evidence as consistent with these predictions and concludes as follows (p. 213):

> The final impression one is obliged to draw from the currently available results is that they provide no really convincing evidence against even the extreme (hubris) hypothesis that all markets are operating perfectly efficiently and that individual bidders occasionally make mistakes. Bidders may indicate by their actions a belief in the existence of takeover gains, but systematic studies have provided little to show that such beliefs are well founded.

While Roll is careful to explain how difficult it is to evaluate these studies, it seems clear that bidding firms are making very little money (if any) through takeovers. Again, version 2 of the winner's curse seems consistent with these data.[9]

Commentary

If my reading of the literature on oil leases and takeovers is correct, namely that the winner's curse is present in these markets, how surprised should economists be? What challenge does the existence of the winner's curse pose to the economists' paradigm? McAfee and McMillan (1987, p. 721) in their survey of auctions and bidding say: "Statements about the winner's curse [such as the quote by Dessauer about book publishing that appears above] come close to asserting that bidders are repeatedly surprised by the outcomes of auctions, which would violate basic notions of rationality." I think that what they are saying can be paraphrased as follows: "These statements about the winner's curse suggest that bidders make systematic errors. Economic theory precludes such errors. Therefore, the statements must be wrong." The internal consistency of this position is questionable. I think it is important to keep in mind that rationality is an assumption in economics, not a demonstrated fact. Given the results of the experimental studies, isn't it possible that some bidders make mistakes in these auctions?

---

[8] Of course, it is an article of faith that the stock prices are rational. Indeed, Miller (1977) argues that stock prices in general are afflicted by the winner's curse because the investors most optimistic about a stock will be the ones who own it. Thus, in the case of mergers, the investors most optimistic about the prospects of the merger will end up owning the shares of the acquiring firm. This argument must be tempered by the possibility that pessimists can sell the acquiring firm short. Nevertheless, it is true that many investors, both individual and institutional, do not sell short, and it remains an empirical question whether the existing quantity of short selling is sufficient to prevent a winner's curse from emerging.

[9] For additional information on takeovers see the papers in this issue's symposium.

It is also interesting to note a peculiar tendency among many economic theorists. A theorist will sweat long and hard on a problem, finally achieving a new insight previously unknown to economists. The theorist then assumes that the agents in a theoretical model act as if they also understood this new insight. In assuming that the agents in the economy intuitively grasp what it took so long to work out, the theorist is either showing uncharacteristic modesty and generosity, or is guilty of ascribing too much rationality to the agents in his model. As Kenneth Arrow (1986, p. 391) has said: "We have the curious situation that scientific analysis imputes scientific behavior to its subjects. This need not be a contradiction, but it does seem to lead to an infinite regress."

The possibility of suboptimal behavior by other participants in an auction raises an issue rarely discussed in economic theory, namely what to do when you realize that your competitors are making mistakes. Theoretical treatments of bidding typically assume that bidders are rational and that the rationality of other bidders is common knowledge.[10] Suppose you are Capen and his colleagues and you have figured out the winner's curse. You now have an advantage over other oil firms. How can you exploit your new competitive advantage? If you react by optimally reducing your bids, then you will avoid paying too much for leases, but you will also win very few auctions. In fact, you may decide not to bid at all! Unless you want to switch businesses, this solution is obviously unsatisfactory. You could let your competitors win all the auctions and try to make money by selling their shares short, but this strategy can be risky. In the oil drilling case, for instance, the price of oil skyrocketed, and the price of oil stocks went up too. A better solution may be to share your new knowledge with your competitors, urging them to reduce their bids as well.[11] If they believe your analysis, then the game can be profitable for the bidders. This, of course, is exactly what Capen, Clapp and Campbell did. More generally, the study of optimal strategy for games in which one's opponents are less than fully rational deserves greater attention from economists.

Even once one is aware of the winner's curse, it is easy to fail to appreciate some of the subtle ways in which it can operate. For example, Harrison and March (1984) discuss the concept of postdecision surprise, a situation similar to the second version of the winner's curse in which decision makers systematically observe outcomes which are worse than expected. They show that postdecision surprise will occur for any decision with great uncertainty and/or many alternatives. Thus, the following should be true: for a department recruiting a new assistant professor, the more candidates it interviews at the annual meetings, the better will be the candidate it hires, and the more likely that candidate will be to fail to live up to the department's expectations. Similarly, Brown (1974) discusses the case of capital investment projects within a firm. If many such projects are considered, and only a few are selected, then actual net revenues will tend to be less than projected, even if the projections are unbiased for the complete set of projects considered.

---

[10] Bidding theory is explored in Wilson (1977), Milgrom and Weber (1982) and the recent survey by McAfee and McMillan (1987).

[11] Thanks to Julia Grant for making this point.

The winner's curse in a prototype for the type of problem that is amenable to investigation using modern behavioral economics, a combination of cognitive psychology and microeconomics. The key ingredient is the existence of a cognitive illusion, a mental task that induces a substantial majority of subjects to make a systematic error. The existence of the cognitive illusion was recognized by Capen, Clapp and Campbell and demonstrated by Bazerman and Samuelson, and by Kagel and Levin. Whenever such an illusion can be demonstrated, the possibility that market outcomes will diverge from the predictions of economic theory is present.

I will close with a fortune cookie provided by Capen, *et. al.* "He who bids on a parcel what he thinks it is worth, will, in the long run, be taken for a cleaning."

■ *The background research for this article was conducted by Julia Grant. Helpful comments were provided by Rob Porter, Carl Shapiro, and Hal Varian. Financial support was provided by the Sloan Foundation.*

References

Arrow, Kenneth A., "Rationality of Self and Others in an Economic System," *Journal of Business*, October 1986 *59*, S385–400.

Asquith, P. "Merger Bids, Uncertainty, and Stockholder Returns," *Journal of Financial Economics*, 1983, *11*, 51–83.

Bazerman, Max H. and William F. Samuelson, "I Won the Auction but Don't Want the Prize," *Journal of Conflict Resolution*, December 1983, *27*, 618–634.

Brown, Keith, "A Note on the Apparent Bias of Net Revenue Estimates for Capital Investment Projects," *Journal of Finance*, 1974, *29*, 1215–16.

Capen, E. C., R. V. Clapp, and W. M. Campbell, "Competitive Bidding in High-Risk Situations," *Journal of Petroleum Technology*, June 1971, *23*, 641–653.

Cassing, James, and Richard W. Douglas, "Implications of the Auction Mechanism in Baseball's Free Agent Draft," *Southern Economic Journal*, July 1980, *47*, 110–21.

Cox, James C., and R. M. Isaac, "In Search of the Winner's Curse," *Economic Inquiry*, 1984, *22*, 579–592.

Dessauer, John P., *Book Publishing*, New York: Bowker, 1981.

Dyer, Douglas, John Kagel, and Dan Levin, "The Winner's Curse in Low Price Auctions," unpublished manuscript, University of Houston, 1987.

Gilovich, Thomas, Robert Vallone, and Amos Tversky, "The Hot Hand in Basketball: On the Misperceptions of Random Sequences," *Cognitive Psychology*, 1985, *17*, 295–314.

Harrison, J. R., and J. G. March, "Decision Making and Postdecision Surprises," *Administrative Science Quarterly*, March 1984, 26–42.

Hendricks, Kenneth, Robert H. Porter, and Bryan Boudreau, "Information, Returns, and Bidding Behavior in OCS Auctions: 1954–1969," *Journal of Industrial Economics*, 1987, *35*, 517–542.

Kagel, John H. and Dan Levin, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, December 1986, *76*, 894–920.

Kagel, John H., Dan Levin, and Ronald M. Harstad, "Judgment, Evaluation and Information Procession in Second-Price Common Value Auctions," unpublished manuscript, University of Houston, 1987.

McAfee, R. Preston, and John McMillan, "Auctions and Bidding," *Journal of Economic Literature*, June 1987, 25, 699–738.

Mead, Walter J., Asbjorn Moseidjord, and Philip E. Sorensen, "The Rate of Return Earned by Lessees under Cash Bonus Bidding of OCS Oil and Gas Leases," *The Energy Journal*, 1983, *4*, 37–52.

Milgrom, Paul R., and R. J. Weber, "A Theory of Auctions and Competitive Bidding," *Econometrica*, 1982, *50*, 1089–1122.

Miller, E. M., "Risk, Uncertainty, and Divergence of Opinion," *Journal of Finance*, September 1977, *32*, 1151–68.

Roll, Richard, "The Hubris Hypothesis of Corporate Takeovers," *Journal of Business*, April 1986, 59, 197–216.

Samuelson, William F., and Max H. Bazerman, "The Winner's Curse in Bilateral Negotiations," *Research in Experimental Economics*, 1985, *3*, 105–137.

Weiner, Sheryl, Max Bazerman, and John Carroll, "An Evaluation of Learning in the Bilateral Winner's Curse," unpublished manuscript, Northwestern University, 1987.

Wilson, Robert, "A Bidding Model of Perfect Competition," *Review of Economic Studies*, 1977, *44*, 511–518.

Journal of Economic Perspectives—Volume 2, Number 1—Winter 1988—Pages 203–207

Notes

The **Journal of Economic Perspectives** intends to publish an occasional review of recent papers studying the economics profession. Such papers may be recently published, soon-to-be published, or unpublished. The review would cover such topics as rankings of economics departments for whatever reason and by whatever criteria, studies of the market for hiring economists, trends in publication records, and so on. If you would like your research results to be considered for inclusion in the review, please send a copy of the full paper by April 1 to Professor David C. Colander, Department of Economics, Middlebury College, Middlebury, Vermont 05753.

Call for papers. The **European Economic Association** (EEA) will hold its third annual congress in Bologna on August 27–29, 1988. Contributed papers in all areas of Economics are actively solicited. Submissions should include two copies of the paper and of a one-page abstract, and must be sent before March 1, 1988, to: John Sutton, Chairman, Programme Committee, EEA Conference, London School of Economics, Houghton Street, London WC2A, 2AE, England. The **European Economic Association** is an international scientific body, with membership open to all persons involved or interested in Economics. For further information contact the European Economic Association, Office of the Secretary, 34, voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.

Call for papers. The annual meeting of the **Middle East Economic Association**, an affiliate of AEA, will be held with the ASSA meetings in New York City, December 28–30, 1988. Those interested in having papers considered should send a one-page abstract to Abbas Alnasrawi, Department of Economics, University of Vermont, Burlington, VT 05405, no later than April 30, 1988.

Call for papers. The **Southern Economic Association** will hold its annual meeting at the Hyatt Regency San Antonio, San Antonio, Texas, November 20–22, 1988. The program Chairman is Dr. V. Kerry Smith, Vanderbilt University. For information on submitting papers, please see the Call for Papers in the October issue of the *Southern Economic Journal*, or write to Dr. Joseph M. Jadlow, College of Business Administration, Oklahoma State University, Stillwater, OK 74078.

Call for papers. The editors of ***Advances in Econometrics*** are currently soliciting abstracts for papers in the areas of cointegration, spurious regressions, and unit roots. Papers chosen for publication will appear in the 1989 volume *Advances in Econometrics: Cointegration, Spurious Regressions, and Unit Roots*. Please send by March 1, 1988, to Professor Thomas B. Fomby, Department of Economics, Southern Methodist University, Dallas, TX 75275.

Call for papers. The **Society for the Advancement of Behavioral Economics** will hold its Fourth Annual Conference June 15–17, 1988, in San Diego, CA. Those interested in presenting papers, being discussants, and attending the conference should send an inquiry to Roger Frantz, Department of Economics, San Diego State University, San Diego, CA 92182.

Call for papers. The **Association of Managerial Economists** will hold its sixth annual meeting in New York City, December 28–30, 1988. Three sessions of competitively selected papers will be featured in conjunction with the Allied Social Science Association's (ASSA) Annual Meeting. Theoretical, empirical and policy analyses across a broad spectrum of topics will be featured. Both members and nonmembers are invited to submit papers, offer their services as session chairs and discussants, and/or make program suggestions by July 15, 1988, to Professor Mark Hirschey, AME *Program Chair, Jones Graduate School, Rice University*, P.O. Box 1892, Houston, Texas 77251-1892. Phone: 713-527-8101, ext 2549.

Call for papers. The thirteenth annual conference of the **Association of Private Enterprise Education** will be held in Cleveland, Ohio, April 24–26, 1988. The emphasis of the meeting is on private enterprise, entrepreneurship and economic education. Paper submissions in these areas are invited. Contact W. F. Kiesner, College of Business, Loyola Marymount University, 7101 W. 80th, Los Angeles, California 90045. Phone: 213-642-4569.

Call for papers. The **Association for the Advancement of Policy Research and Development in the Third World** announces its eighth annual conference, "Towards a World Decade for Scientific and Technological Co-operation for International Development," to be held November 20–25, 1988, in Kingston, Jamaica, West Indies. Send one-page paper abstracts, panel proposals and roundtable suggestions by Friday, February 26, 1988, to Dr. Mekki Mtewa, Executive Director, Association for the Advancement of Policy, Research and Development in the Third World, P. O. Box 70257, Washington, D.C. 20024-1534.

Call for papers. ***The Kentucky Journal of Economics and Business*** is seeking papers for publication in its 1988 volume. The *Journal* is an annual, refereed publication that publishes articles of general interest to the economics profession. Style sheets and other particulars are available upon request, but there are no style requirements on initial submissions. The submission deadline for the 1988 publication is April 15. To submit articles send three copies to John J. Bethune, Articles Editor, *KJEB*, Department of Economics, Bellarmine College, Louisville, KY 40205.

Call for papers. The Annual Meeting of the **Association of Environmental and Resource Economists** (AERE) will be held jointly with the AEA in New York, December 28–30, 1988. Those interested in having papers considered for these sessions should send three copies of a one-page abstract to Prof. Charles Howe, Department of Economics 256, University of Colorado, Boulder, CO 80309, no later than June 15, 1988.

The **International Atlantic Economic Conference** announces its twenty-fifth meeting, "A Time for Change," to be held in London, England, April 16–22, 1988. Those wishing to present a paper should include a submission fee of $35.00 for each paper along with two copies of a 500-word summary and a separate cover sheet listing location of conference, name and address and phone number of author(s); institution; number and name of the JEL category of this article. Those wishing to be discussants or chairmen should submit the above information but no submission fee. Send to: John M. Virgo, International Program Chairman, Atlantic Economic Conference, Southern Illinois University at Edwardsville, Campus Box 1101, Edwardsville, IL 62026-1101.

The **Fairchild-Martindale Center at Lehigh University** announces a conference on Economic Aspects of Regional Trading Arrangements to be held May 25–27, 1988, in Bethlehem, PA. For information, contact Conference Services, Rathbone Hall 63, Lehigh University, Bethlehem, PA 18015. Phone: 215-758-3514.

The biennial congress of the **International Joseph A. Schumpeter Society** will be held at the University of Siena, Italy, May 24–27, 1988. The theme of the congress is "The Evolution of Technology and Market Structure in an International Context." For information write Professor Arnold Heertje, 17, Laegiskampweg, NL-Naarden 1412, ER, or Professor Horst Hanusch, Universität Augsburg, Memminger Str. 6–14, D-8900 Augsburg.

The **Census Bureau's Fourth Annual Research Conference** (ARC IV) will be held March 20–23, 1988, at the National Clarion Hotel in Arlington, Virginia. ARC IV will comprise a mix of topics ranging from new survey techniques and census survey automation to measurement problems with foreign trade statistics. For further information contact Maxine Anderson-Brown, ARC Conference Coordinator, Office of the Director, Bureau of the Census, Washington, DC 20233.

The University of Tennessee announces its **International Workshop in Post-Keynesian Economics**, "Problems in Income Distribution: Functional, International, Personal," to be held June 26 through July 3, 1988, in the Great Smoky Mountain Resort area of Tennessee. The program is being organized by Professors P. Davidson and J. A. Kregel. For information, contact Louise Davidson, c/o *JPKE*, University of Tennessee, 508 Stokely Management Center, Knoxville, Tennessee 37996-0550.

The **Economic Science Association** (ESA) was formed in early 1986. The stated purpose of the ESA is "to advance, enhance, or further Economics as an observational science through use of laboratory and field method of observation and data collection under the control and responsibility of the research investigator, and the development of economic theory and statistical or econometric methods based on such direct observations and data." Current plans of the association include joint meetings with the Public Choice Society, March 1988, in San Francisco. For further information write Vernon L. Smith, Department of Economics, University of Arizona, Tucson, AZ 85721.

The **Indian Econometric Society** announces the *Journal of Quantitative Economics* (*JQE*) published biannually by the Delhi School of Economics. The editor is A. L. Nagar. Authors from North and South America who wish to submit their articles to the *JQE* may send them (in triplicate) either to the editorial office of the journal at the Delhi School of Economics, University of Delhi, Delhi 110 007, India, or to Damodar Gujarati at the Department of Economics and Finance, Baruch College (CUNY), 17 Lexington Avenue, New York, N.Y. 10010.

1988–89 Fulbright Awards in Economics. The Council for International Exchange of Scholars has announced that competition for the 1988–89 Fulbright grants in research and university lecturing abroad is in progress. For more information and applications, call or write Council for International Exchange of Scholars, Eleven Dupont Circle N.W., Washington, D.C. 20036-1257. Phone: 202-939-5401.

The **American Statistical Association** (ASA) and the **Bureau of Labor Statistics** (BLS), under a grant from the National Science Foundation (NSF), are conducting a research fellowship program. Fellows and trainees selected by ASA will work at the BLS for periods of up to one year on specific research activities relating to BLS programs. For a description of the program and more information, write to Dr. Fred Leone, Executive Director, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402.

The **Credit Research Center**, an independent nonprofit research center located at the Krannert School of Management at Purdue University, has funded an annual award of $1,500 for research in the area of consumer credit. To be eligible for the award you must submit, by April 1, 1988, a completed research paper on a consumer credit or markets topic. For information, contact Robert W. Johnson, Director, Credit Research Center, Purdue University, West Lafayette, IN 47907.

The **American Council of Learned Societies** offers small travel grants to economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship and/or participation. The deadlines for applications to be received are: for meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45th Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Research in the History of Economic Thought and Methodology, published annually by JAI Press, will inaugurate in 1988 a series of archival supplements. The archival volumes will contain hitherto unpublished materials of varying descriptions. Suggestions and contributions may be sent to Warren J. Samuels, Department of Economics, Michigan State University, East Lansing, MI 48824.

The **National Institute of Mental Health** announces a new research grant program to encourage investigator-initiated research on services to the mentally ill including the priority area of financing such services. For information contact Mr. Paul Widem or Dr. Agnes Rupp, Mental Health Economics Research Program, NIMH, 5600 Fishers Lane, Room 18C-14, Rockville, Maryland 20857. Phone: 301-443-4233.

The Journal of Futures Markets offers discount subscriptions to members of the AEA. Members pay $70.00 for 6 issues (regular rate, $142). Outside the U.S. add $37.00 for air service. Members who wish to take advantage of this special rate may send their orders with payment directly to the Periodicals Division, John Wiley & Sons, Inc., P.O. Box 836, Bound Brook, N.J. 08805, and identify themselves as members of the American Economic Association.

Deaths

William E. Dunkman, professor emeritus of economics, The University of Rochester, July 15, 1987.

Retirements

Karl A. Fox: distinguished professor, Iowa State University Economics Department.

Gian S. Sahota: professor of economics, Vanderbilt University, July 1, 1987.

New Appointments

Richard G. Anderson, Ohio State University: visiting assistant professor, Department of Economics, University of Michigan, Summer 1987.

Frances Antonovitz: associate professor, Department of Economics, Iowa State University, August 21, 1987.

Edward F. Buffie: associate professor of economics, Vanderbilt University, August 1, 1987.

Stephen Blough: instructor, The Johns Hopkins University, July 1, 1987.

Victor Brajer: associate professor of economics, California State Univerity, Fullerton, August 1987.

Donald Cox: associate professor of economics, Boston College, July 1, 1987.

Sanghamitra (Mita) Das, University of Wisconsin-Madison: assistant professor of economics, Indiana University-Bloomington, August 1987.

Peter Eckstein, Michigan State AFL-CIO: visiting scholar, Department of Economics, The University of Michigan, September 1, 1987.

Gerard Gaudet, Université Laval: visiting professor, Department of Economics, The University of Michigan, September 1, 1987.

Peter Gottschalk: professor of economics, Boston College, July 1, 1987.

Jens Haass, Institut fur Theoretische Volkswirtschaftslehre, University of Kiel: visiting scholar, Department of Economics, The University of Michigan, September 1, 1987.

Thomas J. Hoerger: assistant professor of economics, Vanderbilt University, August 1, 1987

Ali M. Khadr, Oxford Institute for Energy Studies: visiting assistant professor, Department of Economics, The University of Michigan, September 1, 1987.

Miles S. Kimball, Harvard University: assistant professor, Department of Economics, The University of Michigan, September 1, 1987.

Wilhelm Kohler, Institut fur Wirtshaftstheorie und Wirtschaftspolitik, University of Innsbruck: visiting scholar department of economics, University of Michigan, September 1, 1987.

Sherrie A. Kossoudji, University of Michigan: assistant professor, The University of Michigan, September 1, 1987.

Cathleen S. Leue, Washington State, Pullman: visiting assistant professor, University of Oregon, fall 1987.

James Levinsohn, Princeton University: assistant professor, Department of Economics, The University of Michigan, September 1, 1987.

E. Scott Mayfield: assistant professor of economics, Boston College, September 1, 1987.

Robert G. McGillivray: senior economist, Niehaus & Associates, Inc., Seattle, Washington.

Bruce Mizrach: assistant professor of economics, Boston College, September 1, 1987.

Giancarlo Moschini: assistant professor, Department of Economics, Iowa State University, August 21, 1987.

Barry Naughton, University of Michigan: visiting assistant professor, Department of Economics, September 1, 1987.

William C. Perkins: assistant professor of economics, Department of Economics and Management, Manhattanville College, August 25, 1987.

Robert W. Rich: assistant professor of economics, Vanderbilt University, August 1, 1987.

Agnes Rupp: economist, Mental Health Economics Research Program, National Insitiute of Mental Health, August 1987.

Joel B. Slemrod, University of Minnesota: associate professor, Department of Economics, Business Administration, The University of Michigan, September 1, 1987.

Timothy M. Smeeding, professor of public policy, George Peabody College: professor of economics, College of Arts and Science, Vanderbilt University, August 1, 1987.

Mark A. Thoma, Washington State, Pullman: assistant professor, University of Oregon, fall 1987.

Weiying Zhang, China's Economist System Reform Research Institute: visiting scholar, Department of Economics, The University of Michigan, September 1, 1987.

Reiner Wolff, University Gesamhochscule Siegen: visiting scholar, Department of Economics, The University of Michigan, September 1, 1987

Resignations

Faye Anderson: Colorado State University, May 20, 1987.

Charles M. Becker: Department of Economics, Vanderbilt University, The Economics Institute, Boulder, CO.

John Haltiwanger: associate professor, The Johns Hopkins University, July 1, 1987.

Lyle Kalish: California State University, Fullerton, June 1987.

V. Kerry Smith: Department of Economics, Vanderbilt University, Department of Economics, North Carolina State University.

To Departmental Representatives and Executive Officers

When submitting information for inclusion in "Notes" please observe the following guidelines:

Calls for papers, notice of professional meetings, and other announcements of interest to economists should be submitted in one paragraph that contains all relevant information.

News of individual members should be labeled as to category: (1) deaths: (2) retirements; (3) foreign scholars visiting the US or Canada; (4) promotions; (5) administrative appointments; (6) new appointments; (7) leaves for special appointments; (8) resignations; (9) miscellaneous. Give individual's name, present place of employment, new title, new institution, and date on which change is to occur.

Deadlines for each issue are: Spring, December 15; Summer, March 15; Fall, June 15; Winter, September 15.

Please send all information to the *Journal of Economic Perspectives*, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544. We reserve the right to edit material received.

AMERICAN ECONOMIC ASSOCIATION

1988 ANNUAL MEMBERSHIP RATES

Membership includes:

– a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

- Regular members with annual incomes of $30,000 or less $38 50
- Regular members with annual incomes above $30,000 but no more than $40,000 $46 20
- Regular members with annual incomes above $40,000 $53.90
- Junior members (available to registered students for three years only) Student status must be certified by your major professor or school registrar $19.25
- In Countries other than the U.S.A., Add $16 00 to cover postage.
- Family members (persons living at the same address as a regular member, *additional memberships without* subscription to the publications of the Association) $7.70

Please enter my subscription for the following period:

☐ **Jan.-Dec.** ☐ **April-March** ☐ **July-June** ☐ **Oct.-Sept.**

First Name and Initial **Last Name** **Suffix**

Address Line 1

Address Line 2

City

State or Country **Zip/Postal Code**

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

☐ ☐

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____________________

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____________________

____

Authorized Signature

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 3–13

Distinguished Lecture on Economics in Government: Thinking About International Economic Coordination

Martin S. Feldstein

I should perhaps begin these remarks by emphasizing that I am not opposed to international cooperation in economic affairs. We as economists recognize that conflict in international trade and the development of protectionist policies would reduce the standard of living around the world. Moreover, the immediate effect of a shift to protectionism could be a major worldwide recession.

The quiet exchange of macroeconomic forecasts and policy plans among government officials that occurs within the framework of the Bank for International Settlements, the OECD, the Group of 7 (the United States, Great Britain, France, Germany, Japan, Italy and Canada) and elsewhere plays an important role in the formulation of domestic economic policies. It is also important for maintaining friendly working relationships among governments that senior officials are not surprised by major policy shifts abroad.

I start with these obvious remarks and with the assertion that I am not opposed to international cooperation in all economic matters because I do not want to be misunderstood when, as in this lecture, I stress the counterproductive consequences of the international coordination of macroeconomic policy.

I do not deny that the economies of the world are linked in a way that makes the monetary and budget policies adopted in one country affect the economic performance of other countries. But I believe that many of the claimed advantages of cooperation and coordination are wrong, that there are substantial risks and disadvantages to the types of coordination that are envisioned, and that an emphasis on international coordination can distract attention from the necessary changes in domestic policy. Moreover, the attempt to pursue coordination in a wide range of macroeconomic policies is likely to result in disagreements and disappointments that

■ *Martin Feldstein is Professor of Economics, Harvard University, and President, National Bureau of Economic Research, both in Cambridge, Massachusetts.*

reduce the prospects for cooperation in those more limited areas of trade, defense and foreign assistance where international cooperation is actually necessary.[1]

In stressing the limited scope for the international coordination of macroeconomic policy and exchange rates, I do not wish to imply that such action is never appropriate. Far from it. There are some small and very interdependent countries where such coordination should undoubtedly be the general rule. There are also some conditions when the potential gains from coordination are such that all countries could expect to benefit from participation. But the active coordination of the macroeconomic policies and of exchange rates among the United States, Japan and Germany will generally be inappropriate. Moreover, as I shall explain in these remarks, the United States is particularly unsuited to participate in an ongoing process of economic coordination.

The Management of Exchange Rates

After less than a decade of floating exchange rates, the sharp rise in the value of the dollar during the early 1980s caused a renewed interest in the possibility of the international management of exchange rates and even of a return to a system of fixed exchange rates. The rise in the real value of the dollar was the primary reason why the United States shifted from a trade surplus in 1981 to a massive trade deficit by 1986. It is not surprising, therefore, that American businesses and their workers in the wide range of industries hurt by the strong dollar called for government action to stop and reverse the dollar's rise.

Until September 1985, the Reagan Administration argued that the value of the dollar should be left to the market and that it was inappropriate to shift U.S. domestic economic policy or to intervene in currency markets in an attempt to alter the dollar's value. This was a continuing source of conflict with the European and Japanese governments that did not like the dollar's rise because of the inflationary pressures that it imparted to their own economies, the rising real interest rates that resulted from the capital outflow to the United States, and the increased support for protectionist trade policies that was developing in the United States because of the surging U.S. trade deficit.

---

[1] My criticisms of international macroeconomic coordination were sketched first in an article that I wrote for the Economist magazine ("The World Economy Today," June 11, 1983). Earlier versions of the current lecture were presented as one of my two 1986 Horowitz Memorial Lectures of the Bank of Israel and as a lecture on the fiftieth anniversary of Nuffield College, Oxford (see Feldstein, forthcoming). The general subject of international economic cooperation (in macroeconomic policy, exchange rate policy, international trade and developing country debt management) was unexplored at an April 1987 conference of the National Bureau of Economic Research. The conference volume presents valuable survey papers, personal comments by individuals who have had significant experience in government and business, and an extensive bibliography that makes it unnecessary to provide such references in the current talk; my own views are summarized in an introductory chapter. The interested reader may also wish to see Feldstein (1987b; forthcoming a, b).

The Plaza Hotel meeting of the finance ministers of the G-5 countries (United States, Great Britain, France, Germany and Japan) in September 1985 was a political watershed in this process. Faced with the reality that the dollar had been declining for more than six months, Treasury Secretary James Baker abandoned the previous Treasury position that the strong dollar was a measure of foreign investors' approval of the economic policies of the United States. He acknowledged publicly that the high value of the dollar was a serious problem for American industry. And, most surprisingly and significantly of all, he agreed to participate in coordinated exchange market intervention aimed at lowering the dollar's value.

Immediately after the Plaza meeting the United States did join with other countries in a major exchange market intervention, selling dollars and buying other currencies. The Japanese central bank also raised short-term interest rates temporarily in order to make yen denominated bonds a more attractive investment and thereby to stimulate the demand for the yen.

Exchange rate targeting and economic policy coordination aimed at achieving desired levels of the exchange rates have been a frequent theme of intergovernmental meetings since that time. The finance ministers of the G-7 countries met at the Louvre in February 1987 to assert their belief in macroeconomic coordination and to call for exchange rate stability. They reaffirmed this call at the June 1987 Venice summit and the September 1987 IMF-World Bank meeting. In more recent days (November and December 1987), European and Japanese government officials have repeated their case for coordinated actions to achieve a stable value of the dollar and a new G-7 communique calling for a stable dollar has been issued.

In contrast to these assertions, I believe that the dollar must continue to decline because the future trade deficit implied by the dollar's current level would be too large to finance otherwise. Experts estimate that without a further decline of the dollar the U.S. trade deficit will remain at more than $100 billion a year and the current account deficit will grow explosively because of the interest and dividends owed on the U.S. net borrowing from abroad. Only a decline of the dollar can achieve the change in the relative prices of American and foreign goods that can induce American consumers to import less and foreign consumers to buy more American-made goods. The more rapid growth in Japan and Europe that the finance ministers continually stress as an alternative to a dollar decline simply cannot be powerful enough to make a significant dent in the U.S. trade deficit. Similarly, while a decline of the American budget deficit would help to shrink the U.S. trade deficit, it would do so by lowering U.S. interest rates which in turn would reduce the value of the dollar; a decline in the budget deficit, while essential for other reasons, is not an alternative to a decline in the dollar as a means of shrinking the trade deficit. (A longer nontechnical discussion of why the dollar must decline is presented in Feldstein, 1987a.)

I believe that the pursuit of exchange rate goals is likely to be both futile and economically damaging, not just in the current circumstances but more generally as well. In the short term, as the experience of October 1987 painfully demonstrated, the expectation that monetary policy would tighten to defend artificial exchange rate levels can destabilize financial markets. The fear that the Fed would push rates even higher than they were in early October—to offset the downward pressure on the dollar that resulted from the unfavorable trade news of October 14th—was one of the key factors that triggered the stock market crash. Over an extended period of time, the primary risk in the pursuit of exchange rate targets is an increase in the rate of inflation in every country. A policy of targeting exchange rates can also hurt the process of capital formation, weaken the capital goods and construction industries, and delay the recovery of the manufacturing industries that are hurt by an overvalued exchange rate.

Nominal and Real Exchange Rates

To understand the likely effects of an exchange rate policy, it is crucial to distinguish between changes in nominal exchange rates and changes in real exchange rates. The change in the nominal exchange rate between two countries is the change in the actual exchange rate that prevails in the market. The change in the real exchange rate is equal to the change in nominal exchange rate adjusted for differences in the inflation rates in the two countries. It is, of course, the real value of the currency that influences the competitiveness of the country's products and thus its exports and imports. This point is worth emphasizing for two reasons.

First, the political discussion about exchange rate targets, target zones and exchange rate management are always in terms of the nominal exchange rates. Therefore even if the politically agreed upon exchange rate targets were achieved and maintained, the relative competitiveness of the countries could change substantially because differences in inflation rates would cause the real exchange rates to change. In short, exchange rate management is likely to be misguided because it focuses on the wrong target.

Second, economic policies that change a country's rate of inflation can easily alter nominal exchange rates without changing the real exchange rates. Therefore, there is a serious danger that an agreement to stabilize exchange rates would lead to increased inflation without any change in the real exchange rates that influence imports and exports.

Consider, for example, the sharp rise in the real value of the dollar that began in 1980 because of the massive increase in current and projected U.S. budget deficits. Reversing the budget deficit could have prevented the dollar's rise or caused it to decline at an earlier time. But on the basis of my experience as a participant in the making of American economic policy at that time, I do not believe that an exchange rate agreement would have produced such a change in budget policy. The Reagan administration was already trying to reduce the budget deficit by a strategy that combined major domestic spending cuts with a conditional tax increase. There is no reason to think that it would have taken a more conciliatory position with Congress in order to deal with the high dollar. Treasury Secretary Regan apparently believed that the budget deficit had no effect on the dollar. In addition, many experts both inside and outside the government argued that the strong dollar showed only that monetary policy was too tight and therefore argued in favor of depressing the dollar by a monetary expansion.

If the United States had agreed in 1983 to stop the dollar's rise, the easiest way would have been for the Federal Reserve to ease monetary policy. The easier monetary policy would have temporarily lowered real interest rates and that in turn would have caused a temporary decline in the real value of the dollar. More importantly, the easier monetary policy would produce inflation and the inflation would cause the dollar's nominal value to decline. In the end, there would have been no change in the real exchange rate or the trade deficit but a higher price level and a high rate of inflation.

Of course, there are two sides to every exchange rate and the relative value of the dollar could have been depressed by a tightening of monetary policy in Europe and Japan that lowered the inflation rates in those countries. There was in fact some tightening of monetary policy in Europe in response to the rising dollar, but that process was inevitably limited by the increased unemployment that resulted from such a monetary contraction. Political reality would inevitably have required that a change in monetary policy to stabilize the dollar be an inflationary easing by the United States rather than a deflationary tightening in the other countries of the world.

The current attempts to slow the rise of the yen and the German mark provide a further example of the risk that exchange rate coordination is likely to be inflationary. The process of intervention in both Germany and Japan has led to very rapid increases in the money supply and in overall liquidity. This had increased the fear of inflation and raised long-term interest rates in both countries. Although the governments of Germany and Japan have asked the United States to raise interest rates as an alternative way of maintaining the value of the dollar, the U.S. has been reluctant to do so in the context of a recovery that has already lasted five years and with an election this year. Although the German and Japanese intervention has not yet produced any significant rise in inflation, a continuation of the associated increases in money and overall liquidity for an extended period would undoubtedly begin to do so.

Gains and Losses of an Overstrong Dollar

There are of course some who believe that the real value of the dollar can be influenced by exchange market intervention alone. Before considering whether such a policy could be expected to work, it is useful to ask whether stabilizing the dollar would in itself have been desirable in today's world economy.

Stopping the rise in the dollar's real value in early 1983 would have reduced the subsequent U.S. trade deficit. Output and employment would have deteriorated less in our export industries and in those firms that compete with imports from abroad. But that is only half of the story. The reduced trade deficit would automatically have meant a reduced capital inflow from abroad. Real interest rates would have been higher in the United States. Investment in plant and equipment and housing construction would have been depressed. The pattern of employment would have been different, but there is no reason to suppose that total employment and output would have been higher. Moreover, the lower rate of capital formation would have reduced the growth of productivity.

In short, stopping the dollar's rise earlier would have had both good and bad effects on the American economy. There is no way to know whether the favorable effects on the export industries and firms that compete with imports would have outweighed the adverse effects on capital investment. I believe that in the absence of clear evidence or analysis to the contrary, it is best to assume that the market produces a better solution than government intervention. That is as true about exchange market intervention as it is about so many other aspects of government interference with the market economy.

The same principle is relevant today. If intervention per se could stop the dollar's fall at the present time, that would contribute to U.S. price stability and would maintain a flow of capital from abroad that dampens real interest rates and provides the funds to finance additional investment in housing and in plant and equipment. But the high dollar would also limit the improvement in the trade deficit, depressing activity in export industries and those that compete with imports from abroad. Ultimately, moreover, the dollar's adjustment would have to occur and would have to be larger because the dollar had contributed to a greater accumulation of net debt from the United States to the rest of the world.

Exchange Market Intervention

Although substantial past experience and economic logic both imply that sterilized exchange market intervention (i.e., currency intervention with no change in domestic monetary policy) cannot be expected to have a sustained effect on exchange rates, the fall of the dollar in the year after the G-5 finance ministers meeting at the Plaza Hotel in September 1985 and the relative stability of the dollar in the seven months after the Louvre meeting in February 1987 have led some government officials and others to attribute significant power to coordinated exchange rate intervention. I believe that conclusion is a misreading of the recent experience.

There is in fact no evidence that the G-5 Plaza meeting or the subsequent coordinated intervention had any lasting effect on the dollar's rate of decline. In the days immediately after the G-5 meeting there was a four percent decline in the dollar's value. But the rate of decline of the dollar during the subsequent year was no greater than it had been in the six months before the G-5 meeting.

Immediately after the Louvre meeting the central banks were able to prevent the dollar's continued decline by a combination of massive intervention and the threat of even more intervention in the future. But after a short time, continued stabilizing of the dollar required a significant rise in U.S. interest rates relative to interest rates in Japan and Germany. The resulting tight monetary policy in the United States and the fear of further Fed tightening in response to market pressure for a lower dollar were important causes of the stock market crash. As soon as the United States declared that monetary policy would no longer be directed at stabilizing the dollar, the dollar's value responded to market pressures and began to decline sharply. The experience in 1987 thus again demonstrates the inability of intervention to shift currency values for more than a very short time.

The Coordination of Macroeconomic Policies

Coordinated exchange rate intervention is only one aspect of the potential coordination of macroeconomic policies. Indeed there are many who accept that exchange rate intervention per se is powerless and who therefore argue for a more general coordination of monetary and budget policies. The Tokyo summit in May 1986 emphasized such macroeconomic coordination and called for a new "multilateral surveillance" procedure managed by the International Monetary Fund.

The first such multilateral surveillance session was held in conjunction with the IMF annual meetings in September 1986. The agenda was predictable. The United States called on Japan and the European countries to expand their economies more rapidly in order to increase their imports from the United States. The Europeans and Japan called on the United States to reduce its budget deficit and to slow the decline of the dollar relative to their currencies, an inherently self-contradictory request. No serious agreements were reached. The Germans explained that their economy was expanding rapidly, that their monetary policy was if anything too loose already, and that a tax cut was scheduled for 1988. The Japanese noted that they were moving gradually toward fundamental changes in their economy and Treasury Secretary Baker said the United States was committed to eliminating the budget deficit over the next five years as promised in the Gramm-Rudman-Hollings legislation. In short, every country said it would go on doing just what it had been doing. And, despite the subsequent pronouncements at the Louvre meeting and the Venice summit, each country continued to pursue what it saw as its own interest, generally unaffected by the "international coordination" process.

Only when the government of Japan became convinced in early 1987 that the sharp rise in the yen-dollar rate during 1986 would create a serious downturn for the Japanese economy did Japan abandon its goal of budget deficit reduction and adopt a package of fiscal and monetary stimulus. In contrast, Germany, despite its high unemployment and zero inflation rate, has not adopted any significant expansionary policies because of a government decision to give primacy to preventing inflation and to reducing domestic labor market rigidities. And the $70 billion decline in the U.S. budget deficit between 1986 and 1987 was not the result of international actions but of domestic politics and the peculiarities of budget accounting. This failure of the summit and other international coordination discussions to achieve any changes in domestic economic policies should not be seen as a surprise or as a reflection of the fact that the process of explicit coordination only began in 1986. It should be seen as the most likely outcome of all such meetings. The experience at and immediately after the September 1987 IMF-World Bank meetings again demonstrated the emptiness of the coordination process and the inclination of major governments to pursue their own self interests.

Constitutional Limits on American Participation

A primary reason why macroeconomic policy coordination cannot work as envisioned by its advocates is that the United States is constitutionally incapable of participating in such a negotiation. The separation of powers in the American form of government means that the Secretary of the Treasury cannot promise to reduce or expand the budget deficit or to change tax rules. This power does not rest with the president or the administration but depends on a legislative agreement between the president and the Congress. In this sense, the United States' participation in any macroeconomic coordination process is fundamentally different from the participation of a country with a parliamentary system in which the prime minister or the prime minister's representative can commit to a change in the nation's economic policy.

It is sometimes suggested that an American Treasury secretary or president, although incapable of promising the enactment of specific budget legislation, could promise that the American government would support a particular legislative proposal. Such a suggestion ignores the negotiating nature of the relation between the president and the Congress. The president cannot propose the legislative package that he expects Congress to enact and he cannot disclose the extent to which he is willing to compromise on specific issues.

Because it is impossible for the United States to reach an international agreement on specific budget, tax or spending policies, any actual agreement would have to call for a change in U.S. monetary policy. International economic coordination could thus be a recipe for the wrong policy change in the United States. For example, if the United States had entered into an agreement in 1983 or 1984 to shift policy in a way that reduced the value of the dollar, the agreement would have required an inflationary increase in the supply of money. At the present time, the finance ministers of other major industrial countries are urging the United States to defend the dollar by a tight monetary policy that could push the U.S. economy into an unnecessary recession. Moreover, even agreements on monetary policy would be beyond the legal authority of the Treasury secretary or the president since it would require the concurrence of the independent Federal Reserve.

Even if the United States could participate effectively in international macroeconomic coordination, I believe that the incentive for the United States and other major industrial countries to do so is more apparent than real. Uncertainties about the actual state of the international economy and uncertainties about the effects of one country's policies on the economies of other countries make it impossible to be confident that coordinated policy shifts would actually be beneficial. The Germans still remember and regret their agreement to serve as a "locomotive" for worldwide expansion after the 1978 Bonn summit meeting. Today the Germans argue that foreign pressure for more expansionary German policy exaggerates the potentially favorable effects of such a change on foreign economies and ignores both the underlying strength of German demand and the need to achieve structural reforms in the German labor market as a precondition for reduced unemployment. A few years ago U.S. Treasury Secretary Regan argued that foreign pressure for the United States to reduce its budget deficit was misplaced because budget deficits do not influence interest rates, exchange rates and capital flows. When there are fundamental disagreements about the way the world economy works, there is little reason to believe that coordinated policy will produce improved performance.

It is important to remember that to a very great extent a nation has the ability to achieve its economic goals by itself. Although the levels of demand, inflation, and interest rates in one country do affect the economies elsewhere, a country can manage its own monetary, budget and tax policies to offset many of the potential influences from abroad. Thus changes in domestic monetary policy can counteract the contractionary or inflationary pressures from abroad and structural tax policies designed to encourage investment can offset the effects of a worldwide rise in real interest rates.

There is also a serious risk that economic summits and ministerial meetings will inhibit appropriate changes in national economic policies. Governments may not take the politically painful steps that they should because they believe that foreign actions will make such policies unnecessary or because they want to use their own lack of action as part of a bargaining strategy to induce desired policies on the part of foreign governments. Thus the assertion at the Louvre ministerial and at the Venice summit that the dollar would stop declining may have contributed to a complacency that reduced pressure on the German government to provide adequate domestic stimulus and that may even have contributed to the failure of the American Congress to make greater progress in reducing future budget deficits.

Concluding Remarks

It is unfortunately easy and often politically convenient to exaggerate the potential gains from international economic coordination and to understate the ability of a nation to guide its own economic future. Our politicians and those of other leading countries should not be allowed to escape their responsibilities by blaming poor domestic economic performance on the policies pursued abroad. Similarly, it would be a serious mistake if the pursuit of international coordination in exchange markets and in macroeconomic policy management became an excuse for not pursuing appropriate domestic policies.

Washington's explicit recognition of its responsibility for America's economic future would also reassure a nation that has become unnecessarily frightened by the prospect that international economic coordination will collapse. Unfortunately, ever since the May 1986 summit the U.S. administration and the governments of the other industrial countries have emphatically asserted that international economic cooperation is crucial to a healthy international economy in general and to continued U.S. growth in particular. Since such assertions are not justified by the extent of the actual interdependence of the industrialized nations, Americans have been inappropriately worried about whether policy coordination would continue.

Because foreign governments will inevitably pursue the policies that they believe are in their own best interests, it was inevitable that the process of international coordination would eventually collapse. This began with the actions of the German and Japanese governments in early October and was underscored by the statements of Secretary Baker just before the market crash. But the real problem contributing to the market decline was not the collapse of international macroeconomic coordination per se, but the false impression created by governments that healthy expansion requires such international coordination. Moreover, what appeared as cooperation by the European governments in reducing interest rates since late October was motivated by their desire to offset the contractionary effects of the worldwide stock market declines and the falling dollar, rather than as a gesture of cooperation with the United States.

The United States should now explicitly but amicably abandon the policy of international coordination of macroeconomic policy. We should continue to cooperate with other countries by exchanging information about current and future policy decisions, but we should recognize explicitly that Japan and Germany have the right to pursue the monetary and fiscal policies that they believe are in their own best interests.

It is frightening to the American public and upsetting to our financial markets to believe that the fate of our economy depends on the decisions made in Bonn and Tokyo. Portfolio investors, business managers and the public in general need to be reassured that we are not hostages to foreign economic policies, that the United States is the master of its own economic destiny, and that our government can and will do what is needed to maintain healthy economic growth.

Although international coordination of macroeconomic policy-making sounds like a way to improve international relations more generally, there is a serious risk that it will have the opposite effect. An emphasis on international interdependence instead of sound domestic policies makes foreign governments the natural scapegoats for any poor economic performance. Pressing a foreign government to alter its domestic economic policies is itself a source of friction and the making of unkeepable promises can only lead to resentment. It would in general be far better if the major industrial countries concentrated on the pursuit of sound domestic economic policies and reserved the pursuit of international cooperation for those subjects like international trade and national security in which cooperation is truly essential.

■ *This article is a condensed version of the 1987 Distinguished Lecture on Economics in Government, Joint Session of the American Economic Association and the Society of Government Economists, Chicago, December 28, 1987. A copy of the full lecture is available from the author.*

References

Feldstein, Martin, "Correcting the Trade Deficit," *Foreign Affairs*, Spring 1987a, 795–806.

Feldstein, Martin, *International Economic Cooperation: Summary Report*, National Bureau of Economic Research, 1987b.

Feldstein, Martin, "Rethinking International Economic Cooperation," Oxford Economic Papers, forthcoming.

Feldstein, Martin, *International Economic Cooperation*, Chicago: University of Chicago Press.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 15–40

The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth

Franco Modigliani

The purpose of this paper is to review what economists know at present about the following question: How large a portion of the existing wealth is the result of a bequest motive, that is, of accumulation for the specific purpose of leaving bequests? I will also endeavor to clarify why an answer to this question is of interest.

In the early Keynesian period when the study of national saving first attracted wide interest, relatively little attention was paid to what led people to save, though it was generally understood that the main systematic reason was to leave bequests. J. M. Keynes, in the famous chapter 9 of the *General Theory of Employment, Interest and Money* (1936), had listed seven distinct motives for saving besides the leaving of bequests. But five of these, which include "to increase one's future income," or "to insure one's independence and power," implied that all, or nearly all, the accumulation would finally wind up as bequeathed wealth. This, in turn, meant that most private wealth originated through bequests—that is, it either had been received through bequests or was destined to be bequeathed. And, indeed, how else could society accumulate wealth?

However, Keynes (1936, p. 107) also mentioned two further motives: "Precaution," that is, "to build up a reserve against unforeseen contingencies;" and "Foresight," that is, "to provide for an anticipated future relation between the income and the needs of the individual or his family different from that which exists in the present, as, for example, in relation to old age, family education, or the maintenance of dependents." These motives, in contrast to most of the previous ones, have the

---

■ *Franco Modigliani is Institute Professor and Professor of Economics and Finance, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts and the 1985 Nobel Laureate in Economic Sciences.*

characteristic that the current saving is motivated by and destined to pay for future consumption through later dissaving.

But Keynes himself did not dwell on the implications or importance of this source of saving and wealth. This task was first partially undertaken by Harrod (1948) who referred to the transitory accumulation as "hump saving," although it would be more correct to call it "hump wealth." The Life Cycle Model (Modigliani and Brumberg, 1954; 1980) can be seen as an endeavor to study the magnitude and implications of transitory saving and hump wealth by relating it to the classical theory of consumer choice and more particularly to the hypothesis of optimal allocation over time, elaborated by Irving Fisher (1930). According to this hypothesis, the planned consumption path reflects the allocation of life resources to consumption over the life span. At least in the absence of bequests, this implies that there will tend to be saving resulting in transitory accumulation of wealth when current income is above, and/or current consumption below, average. There will be dissaving, financed from the transitory accumulation, in the opposite case.

One of the most significant early results of the Life Cycle Hypothesis was to establish that, even in the absence of bequests, the mere fact that income dries up with retirement could generate, for the entire economy, an amount of (hump) wealth quite large relative to income. Specifically, assuming a stylized life cycle of income and consumption—to wit, level consumption through life, income constant up to retirement and a stationary economy—it was shown that the ratio of wealth to income would be equal to one-half the length of retirement which (at the time) could be taken as of the order of 5. The wealth estimate of Goldsmith, which became available around that time, showed that for the United States the ratio of private net worth to disposable income was of that magnitude, or, if anything, a little lower. It was thus at least conceivable that the bulk of wealth might be acquired not by intergenerational transfers but instead be accumulated from scratch by each generation, to be consumed eventually by the end of life.

Still, the existence of some role for bequeathed wealth, especially among those in the upper strata of the distribution of wealth (human and nonhuman), was self-evident. For this reason, in the second of the two papers laying the foundation of the LCH (Modigliani and Brumberg, 1980) and several later ones (for example, Modigliani and Ando, 1957; Modigliani, 1975), the question of inheritance was dealt with explicitly. It was shown that bequests could be readily incorporated in the model without changing its basic implications, provided the leaving of bequests satisfied two reasonable assumptions.

First, the share of its resources that a household earmarks, on the average, for bequests is a (non-decreasing) stable function of the size of its life resources *relative* to the average level of resources of its age cohort.

Second, the frequency distribution of the ratio of life resources to mean life resources for each age group is also stable in time.

While the second assumption is hard to test, the first has recently received strong support from a study by Menchik and David (1983) described below. If the bequest motive satisfies these assumptions then one can readily establish that the ratio of inherited wealth to income will tend to be constant and independent of per capita income (Modigliani, 1975). But in turn, this means that a generalized life cycle model including both hump and inherited wealth will continue to exhibit all the basic macro properties of the elementary model, including 1) the saving rate is independent of per capita income, but rises with its rate of growth, and, 2) aggregate saving is not merely a reflection of individual thriftiness in the sense that a country with higher growth will save more than another, even though individuals in each country have identical life cycle saving behavior.

But acknowledging that aggregate wealth could arise from both transitory hump wealth and from the transfer of wealth through bequests from one generation to the next conveys nothing about the importance of each of these processes in accounting for existing wealth. In particular, is the bequest motive the main source of existing wealth, as supposed by the traditional view, or is it swamped by hump saving?

This question has attracted attention at least since the early 1960s when a number of investigations (reviewed below) were undertaken. The interest was spurred, in part, by scientific curiosity. But the question also has relevance for the design of economic policies because the two sources of wealth may be expected to respond to very different stimuli. According to the Life Cycle Hypothesis, hump wealth should respond to variables or institutions like length of retirement, family size, liquidity constraints, uncertainty of income (at least from labor), private and public pension arrangements, and health insurance. Most of these variables would likely have little effect on bequests, though, admittedly, economists know rather little as to what, other than estate and gift taxation, would have a significant impact on bequests.

Thus, knowledge of the relative contributions is important to assess the effectiveness of measures designed to affect saving and wealth as well as the effects on wealth of measures intended to achieve other goals, such as estate taxation designed to reduce economic inequalities.

Estimates of the Share of Inherited Wealth

From the early studies of the 1960s and until the recent contribution of Kotlikoff and Summers, it has been generally accepted that the importance of the contribution of the bequest process to total wealth could be measured by the ratio of wealth received through inheritance and major gifts by those living to total private (nonhuman) wealth. Though this ratio may not be an altogether appropriate measure of the importance of the bequest motive (as discussed below), I will start by reviewing available estimates of the share, both because most of the existing information relates directly to this variable and because this measure provides a useful building block for alternative measures.

Several methods have been utilized to estimate the share of private wealth accounted for by bequeathed wealth: 1) asking people directly through a survey; 2) estimating the annual flow of bequests and then using an appropriate "blow up" factor to infer the *stock* of inherited wealth; and 3) inferring inherited wealth indirectly by first estimating life cycle wealth and then subtracting it from an independent estimate of private wealth.

The Survey Method

This was the first method used, in connection with three U.S. studies carried out in the 1960s. In the study by Morgan et al. (1962), respondents were asked "about your own personal reasons for saving," with twelve answers offered. Only 3 percent mentioned "to provide an estate for family"! All the reasons most frequently mentioned are consistent with hump saving, such as "old age" (41 percent), "emergencies" (32 percent), and "children's education" (29 percent). The proportion of people referring to the bequest motive does increase, as one might expect, with wealth, but even in the top wealth class (half a million 1963 dollars and over), only one-third of the respondents mention it. Very similar results were found in the 1964 Brookings Study (Barlow et al., 1966).

The results of the three early surveys are summarized in the first four rows of Table 1. The estimates of the share of presently held wealth resulting from bequests (and possible major gifts) is reported in the last column. In view of the nature of the studies, these estimates are neither very precise, nor entirely comparable; for example, the study in Row 2 covered only people in the top 10 percent of the income distribution. Yet the small scatter of results appears to support the view of a fairly modest ratio of inherited to total wealth, between one-tenth and one-fifth (Modigliani, 1975). All the studies report that the share rises with the income and wealth of the recipient.

However, these figures may incur some suspicion because the respondents' replies were largely undocumented and could suffer from serious recall biases. For instance, it is not inconceivable that respondents would tend to underestimate, systematically and significantly, the extent to which their wealth was bestowed on them by others, rather than representing the fruits of their own efforts. It is therefore useful to compare these direct estimates with those obtained through alternative methods.

The Flow of Bequest Approach

This method starts from an estimate of the flow of bequests, which is then translated into a stock of inherited wealth. Three alternative methods have been employed to estimate this flow.

Estimates Based on Age-Specific Wealth Holding and Mortality. This method, first applied by Kotlikoff and Summers (1981), infers the flow of bequests from the distribution of wealth by age and age-specific mortality rates. They apply this method for the year 1962, relying on the distribution of wealth by age provided in Projector and Weiss (1964). From this source they arrive at an estimate of the transfer flow of $11.9 billion. To this, however, one must add bequests received through life insurance

Table 1

Estimates of the Share of Wealth Resulting from Transfers

Source	*Nature of data*	*Nature of transfers included*	*Valuation*	*Share of wealth from transfers (%)*
Morgan *et al.*, (1962)	response to survey question on size of transfer received	all(?)	at time of receipt(?)	less than 10
Projector & Weiss (1964)	response to survey question on share of wealth from transfer	inheritance and large gifts from outside family	at time of receipt	[16][a]
Burlow et al., (1966)	response to survey question on fraction of wealth transfers. Limited to income $10,000 and over	inheritance and gifts	at time of receipt	one-seventh
same	same	same	at time of receipt, plus capital appreciation to present	less than one-fifth
Kotlikoff & Summers (1981)	estimated from wealth of those dying: age gap: 25–30 years	intergenerational bequests	at time of receipt, corrected for price level changes	17
Projector & Weiss (1964)	response to question on transfer received during year; age gap: 25–30 years	inheritance and gifts from outside family	same	$15\frac{1}{2}$
Menchik and David (1983)	Probate records; age gap: 25 years	all bequests other than intraspousal	same	$18\frac{1}{2}$

[a] Wealth-weighted average of respondents answering that inherited assets were a "substantial portion" of total assets.

death benefits and newly established trusts (since neither of these items is included in the data), which raises the estimated intergenerational flow to $16 billion.

To transform the flow into a stock of wealth, Kotlikoff and Summers rely on the assumption that beneficiaries, on average, receive a constant fraction of their life labor income in the form of bequests, and that the average gap between the age of the bequeathers and that of the beneficiaries is a constant, say g years. These assumptions insure that, in steady growth, the ratio of inherited to total wealth tends to a constant. Denote by B the current flow of bequests left (and received). Then, if the economy were stationary, one can readily infer that the aggregate amount of bequests received by those currently living, say T, would come to B for each of g years, or $T = gB$. However, if population and/or productivity increase at a stable rate, say n, then the flow of transfers made and received t years earlier would be smaller than B, amounting to Be^{-nt}. One can then infer that the stock of inherited wealth will be

$$T = \frac{(1 - e^{-ng})}{n} B. \tag{1}$$

(Note that this equation is not the one used by Kotlikoff and Summers. The reasons are discussed below.) To arrive at an estimate of the aggregate transfer T received by all those living, one still needs values for the growth rate n and the age gap g. For the growth rate n, one can use the estimate used by Kotlikoff and Summers (1980), namely 3.5 percent. For the age gap, on the other hand, there is very little solid information to rely on. I will follow their choice of 30 years, though 25 years might be a more reasonable guess (as suggested by data for the United Kingdom assembled by the Royal Commission on the Distribution of Income and Wealth).

With these parameters, the stock of inherited wealth T implied by equation (1) is 18.6 times the flow (as compared with 30, in the absence of growth). With the annual flow of bequests B estimated at $16 billion per year, the stock of inherited wealth T comes to just below $300 billion. Now, in 1962 the stock of household net worth, as estimated by the Federal Reserve Board (1981), came to $1.75 trillion. Thus, the flow of bequests is somewhat short of 1 percent of wealth, and the estimated shares of wealth resulting from bequests can be placed at 17 percent (as reported in Table 1, Row 5). This figure is broadly consistent with the various estimates based on the direct survey method summarized in Table 1 (though it should be recognized that the Kotlikoff and Summers measure of transfers is probably somewhat less inclusive than that used in the surveys, since they purport to estimate "distant in age" intergenerational transfers).

Estimates Based on Survey Information. In the Projector (1968) follow-up study, respondents were asked: "During 1963 did you... receive any gifts or inheritances from persons outside the family?" The answer to this question should provide an alternative estimate of the flow of gifts and bequests. A recent unpublished tabulation of this data yields an estimate of the average reported amount received of $205 per household, or a total of $11.6 billion,[1] compared to the Kotlikoff and Summers estimate of $12 billion before corrections. After correcting for trusts, one arrives at a share of inherited wealth for the bench mark year 1962 of some 15½ percent (Table 1, row 6), again well within the range of the earlier estimates.

Estimates Based on Probate Statistics. The third approach, which is potentially the most promising in terms of the objectivity and quality of the information, is becoming feasible at present through the painstaking efforts of Menchik and David. Their 1983 paper relying on probate records of people who died in Wisconsin between 1947 and 1978 provides information from which one can estimate that the mean bequest (including life insurance proceeds and reported inter vivos transfers) for all male decedents is \$20,000 (in 1967 dollars).[2] Accepting this figure as representative of both sexes (though probably upward-biased), and multiplying by the 1.5 million adults deceased in 1962, the (upper) estimate of the total flow of bequests in that year would be \$29$\frac{1}{2}$ billion (in 1967 dollars). But one should subtract transfers between husband and wife from this figure.

---

[1] I wish to express my gratitude to Kim Kowalewski of the Federal Reserve Bank of Cleveland for computing these averages from an edited version of the tape of the "Survey of Changes in Consumer Finances" (Projector, 1968).

On the basis of estimates of interspousal transfers provided by David and Menchik (1982, Table 5) and vital statistics, I have concluded that the amount of such transfers can be placed at just above \$8 billion (Modigliani, 1984) yielding an estimate for the overall non-intraspousal transfer flow of \$19$\frac{1}{2}$ billion, in 1962 prices. This estimate is substantially above that of \$16 billion arrived at by Kotlikoff and Summers, but this result can be explained at least partly by the consideration that their measure is meant to include only intergenerational transfers.

Data presented by David and Menchik (1982) suggest that intergenerational transfers may account for around 60 percent of the total transfers. If so, the Kotlikoff and Summers estimate of intergenerational transfers would appear larger than that implied by the probate statistics, though the difference does not appear worrisome in view of the many guesses involved in each estimate.

In converting the estimate of the annual flow of total non-intraspousal transfer of \$19$\frac{1}{2}$ billion into a share of wealth, one must remember that the average age gap between bequeather and beneficiary must presumably be appreciably lower for this flow than for purely intergenerational transfers, though it is hard to say by how much. If the age gap is, say, 15 years, for the roughly one-third of bequests that are not intergenerational, while the average for the remaining two-thirds that are intergenerational stays at 30 years, then the average gap would be 25 years, implying a "blow up" factor of 16.7 and an inherited share of wealth of some 18$\frac{1}{2}$ percent (see Table 1, Row 8).

In summary, direct estimates of the annual flow of bequests based on three different approaches appear broadly consistent, especially if one allows for some differences in the definition of the flow being measured. They all imply a rather small annual *flow compared with total wealth*, around 1 percent, more or less. It is this modest annual flow that insures that the total stock of bequeathed wealth (the sum of the annual flow over the transfer gap, adjusted for the effect of growth), is itself modest, between 15$\frac{1}{2}$ and 18$\frac{1}{2}$ percent, the latter measure corresponding to the broadest definition of the bequest flow. These estimates are clearly not inconsistent with the figures suggested by the three direct surveys.

---

[2] This figure, as well as all averages cited below, represents the mean for all persons dying in the course of the three decades spanning the years 1947 to 1978. Presumably, in the course of this span of time, bequests have tended to rise with the rise in per capita income. However, since this discussion is using 1962 as a benchmark, a year which is right in the middle of the period covered, it is reasonable to assume that the mean for the entire period is a fair approximation to the mean for the years of the early 1960s.

Indirect Estimates Based on Life Cycle Saving and Wealth

The essence of this approach is that it endeavors to estimate hump saving and wealth directly, and then derives transfer wealth as a residual. Following this method, Ando and Kennickell (1985) have set out to estimate life cycle wealth by starting from available annual estimates of national saving. For each year, they allocate the aggregate wealth over the age groups present, using the savings-age profiles derived from survey data (in their case, the Bureau of Labor Statistics' Consumer Expenditure Survey for 1972 and 1973) to obtain estimates of saving for individual age group cohorts in any given year. By summing up the saving of any given cohort up to any given year, they can estimate the self-accumulated, or non-inherited, wealth of that cohort in that year (except for capital gains and losses, discussed below). Finally, they arrive at national non-bequeathed wealth in any given year by summing over the cohorts present in that year. The result of this calculation, using both the 1972 and the 1973 age profiles, is reported for every year from 1960 to 1980, and then compared with the actual value of household net worth for the last quarter of each year (Federal Reserve Board, 1984).

The shares of self-accumulated wealth implied by the two profiles are very similar except that the one for 1973 is consistently five percentage points higher. Based on the 1973 profile, one finds that from 1974 to the end of the series in 1980, the share of self-accumulated wealth falls between 80 and 85 percent, a finding remarkably consistent with the consensus of the studies already described. For the earlier years, the share of self-accumulated wealth is smaller, around 60 percent until 1968, then drifting up to over 70 percent by 1973. But the lower figure for the early years may, at least partly, reflect a downward bias in the Ando and Kennickell estimate arising from the fact that their estimate of self-accumulated wealth omits change in real wealth arising from capital gains or losses. In the period before 1974, capital gains were unquestionably significantly positive, and hence self-accumulated wealth is underestimated. On the other hand, from 1974 to 1980, this effect was, presumably, undone by the depressed state of the stock market, even though this may have been partly offset by rising real estate values.

Thus, this alternative and totally independent method yields results which for the last decade are broadly similar to those produced by all other methods of Table 1.

Estimates from Other Countries

It would be interesting to compare these findings with estimates for other times and countries. The only relevant information of which I am presently aware relates to the United Kingdom and is the result of the work of the above-mentioned Royal Commission on the Distribution of Income and Wealth. Their method relies on an estimate of the age pattern of recipients of bequests left by decedents of different age and sex, and on information on the flow of bequests for a long stretch of years terminating in 1973. From this information, they obtain an estimate of the 1973 stock of inherited wealth. Combining this data with total wealth in that year (based on the estate duty method) they arrive at an estimate that inherited wealth is 20.3 percent of total wealth, a share which rises to 24.7 percent when gifts ("all forms of transmitted wealth") are included (Report No. 5, Chapt. 9, tables 90 and 91). This figure is of the same order as those found for the United States, but it includes interspousal transfers. Considering that interspousal transfers in the United States seem to represent somewhat over one-fourth of the total, the share of total wealth that was inherited in the United Kingdom appears to be, if anything, a little smaller than suggested by U.S. data.

Evidence Suggesting a Major Role for Bequests

The evidence presented in the previous section suggests that the bequest process plays an important, but quantitatively modest, role in the process of accumulation of national wealth. However, this conclusion has been seriously criticized and challenged on the basis of variety of evidence seemingly inconsistent with this conclusion.

The Behavior of Wealth in Old Age

In the stylized, pure life cycle model, wealth must be clearly declining after retirement, and at a sufficiently fast pace to reach exhaustion at the end of life. The actual behavior of wealth by age seems quite different, especially after correcting for the fact that successively older households belong to cohorts which typically enjoyed a smaller life income. Several studies find that dissaving in old age is small at best (Fisher, 1950; Lydall, 1955; Menchik and David, 1983). Some studies (for example, Mirer, 1979) even find that wealth actually continues to rise in retirement. Such a finding, if valid, would certainly be inconsistent with significant smoothing of consumption over life; it would mean that as income dries up with retirement, the reduction is entirely absorbed by consumption and accumulation can continue. Thus, even with a bequest motive, smoothing implies that wealth must decline after retirement unless retired consumption can be entirely financed by the return on accumulated wealth (Hurd, 1986), an outcome which would require a ratio of wealth at retirement to per capita consumption far larger than the ratios typically observed.

Actually, most recent analysts have concluded that the wealth of a given cohort tends to decline after reaching its peak in the age range 60–65 or somewhat beyond it, and in any event after retirement, though the extent of the decline depends on the concept of saving and wealth used (Shorrocks, 1975; King and Dicks-Mireaux, 1982; Avery, Elliehausen, Canner and Gustafson, 1984; Bernheim, 1984; Diamond and Hausman, 1985; Ando, 1985; Ando and Kennickell, 1985; Hubbard, 1986; Hurd, 1986). If one makes appropriate allowance for participation in pension funds, then the dissaving (or the decline in wealth) of the old tends to be more apparent, and it becomes quite pronounced if one includes an estimate of social security benefits and wealth,[3] and if one focuses on retired households. Hurd (1986), relying on a large

[3] Bernheim (1984) finds that including pension and Social Security does not increase the dissaving of the old, but this is only because of his unconventional and questionable valuation of annuities. Given that saving—or income minus consumption—is the change in wealth, the consumption of a pension must reduce wealth.

sample of panel data of retired people over a ten-year span, found that even market wealth declines at appreciable rates of about $1\frac{1}{2}$ percent per year for all marketable wealth, and nearly 3 percent if owner-occupied houses are excluded.[4]

Also, several factors tend to bias the age profile of wealth upward. One source of bias, to which Shorrocks has called attention, arises from the well-known positive association between longevity and (relative) income. This effect means that the average wealth of successively older age classes is the wealth of households with higher and higher life resources, hence the age profile of wealth is upward-biased. In a similar vein, Ando and Kennickell (1985) have found evidence that aged households which are poor tend to double up with younger households and disappear from the sampled population so that the wealth of those remaining independent is again an upward biased estimate of average wealth.

But even allowing for these biases, the rate at which marketable wealth is being drawn down during retirement does not appear consistent with the elementary no-bequest form of the life cycle model. However, this result is not very revealing since the issue is not whether bequests exist, but rather their level of quantitative importance. The evidence on behavior of wealth in old age cannot answer that question. To be sure, the post-retirement (past 65) path, particularly at advanced ages, does tell something about the flow of bequests left. However, the *share* of bequest received in total wealth depends on total wealth and hence also on the path of wealth before retirement. If one fixes the path of per capita wealth from its peak, there will still exist an infinity of possible paths from, say, age 20 to 65, and each of these paths implies a different amount of aggregate wealth. The earlier the average path approaches the peak value, the larger will be aggregate wealth and hence the hump wealth component. So, a slow decline from peak is not inconsistent with inherited wealth being a modest share of total wealth.

Actually, there are good grounds for holding that the observed slowly decreasing path of wealth for the United States is fully consistent with the estimate of a bequest share of roughly one-fifth of total wealth. One of the methods of estimation which yielded the one-fifth figure consisted in using the flow of bequests estimated by Kotlikoff and Summers from age specific wealth holdings and mortality. The wealth holdings by age were those reported by the 1962 Federal Reserve Survey of Consumers' Finances, and that data also exhibits the characteristic of slow decline of wealth in old age. Clearly the one-fifth estimate of the share of wealth from bequests, being derived from such a slowly declining path of wealth, must be fully consistent with that path.

---

[4] Bernheim (1984) relies on the same basic data but breaks the period into two five-year periods, 1969–1975, and 1975–1979. He finds that dissaving is fairly sizable for all groups except couples in the second period. But Hurd's data suggest that this result may not be reliable as it reflects an entirely improbable rise of 23 percent between 1977 and 1979 in the wealth of couples. This rise is most unlikely to reflect voluntary accumulation and must be supposed to arise either from noise in the data and/or from unusual capital gains.

Temporal Changes in Retirement Habits and the Saving Rate

Another piece of evidence that has been adduced against a relatively important role for life cycle accumulation is the fact that since the 1930s retirement ages have fallen and life expectancy has risen, spelling a longer average retirement span. This lengthening should have increased the need for accumulation to finance retirement and thus resulted in a higher saving rate. In fact, there is no evidence of a rise in saving. However, the lengthening of retirement was accompanied by another large scale phenomenon: the Social Security revolution. Social Security should tend to reduce saving, offsetting the rise that should result from a longer retirement span induced, at least in part, by the Social Security (Feldstein, 1977; Modigliani & Sterling, 1983). However, the issue of the interaction between Social Security, retirement and saving is a complex one, beyond the scope of this paper.

In addition, the argument of the previous section applies here as well. The importance of inherited wealth cannot be settled by focusing exclusively on retirement behavior, since the rapidity with which hump wealth accumulates during a lifetime is motivated also by considerations other than length of retirement.

Simulation Studies

One method sometimes used to assess the importance of life cycle saving is to rely on simulation techniques: one assumes values for the preference parameters and for the opportunity set and derives life paths of saving and wealth, which are then aggregated to obtain national totals. In fact, this method was used by Modigliani and Brumberg (1954, 1980) to study properties of the life cycle model. But this technique, while useful to suggest possible ranges of outcomes and the responses to changing parameters, cannot settle the empirical issue of the relative importance of hump wealth and bequeathed wealth, because the outcomes are greatly affected by the choice of certain critical parameters which are largely arbitrary. Evans (1984) has demonstrated this point strikingly in his criticism of a frequently cited paper by White (1978). To illustrate, he shows that under plausible assumptions about the economy's growth rate (3.5 percent) and the rate of return (4 percent) by varying the assumed preference parameters, namely the elasticity of temporal substitution and the rate of time preference, between .25 and .5 and 0 and 1 percent respectively, one obtains simulated values of the life cycle rate of saving varying between 2 and up to 11 percent, which is consistent with a lot of room for bequests at one end or very little at the other.

The Kotlikoff and Summers Estimates

In their 1981 paper cited earlier, Kotlikoff and Summers have also reached the conclusion that the share of inheritance and gifts in total nonhuman wealth in the United States is far higher than the one-fifth share indicated by Table 1. In fact, they argue the share may be as high as four-fifths.

This conclusion rests on two alternative estimates of the share of wealth bequeathed or transferred by gifts. The first is based on a variant of the indirect method later employed in the Ando and Kennickell (1985) study discussed earlier. The difference is that instead of allocating saving among groups of households defined by the age of the household head, they impute to each cohort of each sex, in every year since 1900, an income *from labor only* and a consumption. The difference is labelled the cohort's "life cycle" saving which is capitalized and cumulated in order to arrive at an estimate of the "life cycle" wealth of each age and sex cohort and, finally, of aggregate life cycle wealth in a given year. Aggregate inherited (transfer) wealth is obtained by subtracting this estimate from aggregate wealth.

Although they present and discuss several variants of life cycle wealth based on alternative measures of interest rates and treatment of interspousal transfers, I will concentrate on their figure of $733 billion (from their Table 2, LCW2, series 2). This figure is conceptually the most relevant, as it correctly includes the life cycle accumulation of a deceased spouse into the life cycle accumulation of the survivor, instead of including it in inherited wealth. Also, this variant is the one they tend to stress. That figure implies that life cycle wealth is only about 19 percent of total wealth, and that therefore the share of transfer wealth comes to 81 percent!

Their second approach consists in measuring the stock of bequeathed wealth by the method (described earlier) of "blowing up" an estimate of the annual flow of bequests obtained from age specific wealth and mortality rates. Even though they rely on the same method and basic sources, they arrive at an estimate which, though less extreme than 81 percent, is still a good deal larger than ours, namely 46 percent.[5] Why the large discrepancies? I propose to show that they arise primarily from differences between the definitions of "inherited" and "self-accumulated" wealth used by Kotlikoff and Summers and the definitions underlying the estimates reported in Table 1.

Superficially, their definitions do not seem to differ from the usual ones. In fact, they coincide with them in the elementary kind of Modigliani-Brumberg (1980) streamlined model in which the return on capital is zero, all people begin earning at the same constant rate until retirement and there are no bequests. But once these simplifying assumptions are dropped, significant differences came to light.

According to the definition used in Table 1, self-accumulated wealth for an individual household is the summation of saving from the formation of the household to the present, where saving is defined as income (inclusive of capital gains) minus consumption. In turn, aggregate self-accumulated wealth is the sum of self-accumulated wealth over all households present (families and single individuals). Correspondingly, transfer wealth is the sum over all households of bequests (and major gifts) received (in constant prices). This definition differs from that of Kotlikoff and Summers in two ways: the treatment of return on inheritance and the definition of the transfer flow.

---

[5] In their 1981 paper, the figure was actually given as 52 percent, but that was due to an algebraic error in the blow up formula (Modigliani, 1984) which they have since corrected.

First, Kotlikoff and Summers deduct from income and the saving flow, as defined above, the return on inherited wealth, which they add to the flow of bequests. As a result, their inherited wealth is not just the cumulation of bequests received but is instead augmented by the inclusion of the capitalized value of the earnings since receipt of the inheritance. With an average age gap between bequeathers and beneficiary on the order of 25–30 years, this definition adds a great deal to the measure of Table 1.

The second important difference in measuring self-accumulated wealth is that instead of using the household as the basic economic unit responsible for the consumption-saving allocation, Kotlikoff and Summers artificially split the household into individual males and females, to each of whom they impute income and consumption on the basis of their age and family composition. Because of this choice, they are unable to use the formation of the household as the point from which accumulation begins, and are forced instead to pick, more or less arbitrarily, a critical age above which saving is imputed to all members of a cohort, including those still dependent, through an imputation of labor income and consumption. They choose a critical age of 18 for both men and women. This procedure has the implication that the (imputed) consumption of all persons that are above 18 but are still dependent, and hence have no income, is treated implicitly as though it represented a life cycle dissaving; accordingly, it is *subtracted from life cycle saving* and *added to inherited wealth* as those terms were defined in Table 1.

Table 2 provides estimates of the quantitative impact of these definitional differences. I start from the Kotlikoff and Summers measure and show how altering the definitions reduces their estimate until it coincides with the estimates of Table 1.

Effect on the Share Based on the Flow of Bequests. Part A of Table 2 deals with the measure of transfer wealth based on the flow-of-inheritance approach. As already described, their estimate of the share of total wealth that is inherited in this case is 46 percent (see row 1) and mine is 17 percent (see row 4). The adjustment in row 2, amounting to 5 percentage points, is the only one that reflects differences of assumption rather than of definition. Specifically, in reporting the correction for an error in the original "blow up" formula, Kotlikoff (1987) indicates assuming that, on the average, bequests are left by people aged 65, 10 years before death. I believe it is more reasonable to assume that bequests are left at death.

The next adjustment (row 3) subtracts from the stock of inherited wealth as defined by Kotlikoff and Summers that part which represents the capitalized consumption imputed to all dependent persons over 18 years of age. Actually, Kotlikoff and Summers do not have an estimate of that flow but have tried a partial remedy by adding to the estimated flow of inheritance at least one portion of that missing flow that happens to be substantial and is possible to estimate: the flow of expenditure for college education estimated at $4.6 billion (in 1962 dollars) per year. Eliminating this addition, the share of bequests is reduced by another 9 percentage points.

Row 4 corrects for the different treatment of the return on bequests. If the stock of inherited wealth is defined as the sum of inheritances received, then the relation

Table 2

Reconciliation of Kotlikoff and Summers with other estimates

A. Estimates based on flow of bequests

	Correction (percentage points) (1)	*Corrected share of wealth (%) (2)*
Kotlikoff & Summers estimate		46
Assuming transfer at death	−4.8	41.2
Elimination of educational expenses	−9.2	32
Elimination of capitalization of inheritance	−15	17

B. Estimates based on cumulation of life cycle saving

	Correction (percentage points) (1)	*Corrected share of wealth (%) (2)*
Kotlikoff & Summers estimate		81
Error in treatment of durable goods expenditure	−14	67
Elimination of capitalization of inheritances	−31.5	35.5
Correction for expenditure on dependent over 18 and other unspecified sources	−15.5	20

between the flow and the stock of inherited wealth was shown to be given by equation (1); but when it is defined as the sum of the capitalized value of inheritance received, then the appropriate formula becomes:

$$T^* = \frac{e^{(r-n)g} - 1}{r - n} B \qquad (2)$$

where T^* is their definition of inherited wealth. For their estimated age gap g of 30 years and interest rate of 4.5 percent, equation (2) yields a "blow-up" factor of 35, very nearly twice as large as the value of 18.6 implied by equation (1). Accordingly, when their measure of the share is recomputed using the definition underlying Table 1, the estimate drops dramatically, by 15 percentage points, to the point when it coincides with the 17 percent figure described earlier and presented in row 5 of Table 1.

Estimates Based on Capitalized Life Cycle Saving. Table 2B presents a similar reconciliation for their alternative procedure in which the share of transfer wealth is derived by cumulating capitalized "life cycle" saving of each cohort present in 1974.

Row 1 reports the share of 81 percent corresponding to their preferred estimate of life cycle wealth of 19 percent. The first correction in row 2 arises from an error in their calculation of consumption and saving. In the figures they present, they measured consumption as inclusive of the purchase of durable goods, instead of treating such goods as a depreciable investment including only current year estimated depreciation in consumption and the excess of purchases over depreciation as a saving to be cumulated into a stock. Their calculations could, therefore, be expected to produce a large downward bias in the estimate of life cycle wealth, especially since in the United States younger age groups will tend to be significant investors in durables, while older people tend to disinvest.

I have tried to estimate the magnitude of this error, as Kotlikoff and Summers have kindly made available their basic data and helped in carrying out the necessary, fairly extensive computations. The correction was found to increase the estimates of life-cycle wealth, as expected, but by an amount so large—26 percentage points—as to raise questions about its plausibility. However, the fact that such a wild estimate could be generated using Kotlikoff and Summers's method and a set of assumptions which they regarded as reasonable in the context of their approach does raise some question about the reliability of their capitalized life cycle saving method.

Kotlikoff and Summers (1986) have subsequently suggested a different correction for the error in the numerator of their share—which arises from omitting the net accumulation of durables—by making the same error in the denominator, that is, not including durables as wealth. With durables representing some 14 percent of the stock of wealth, they arrive at a correction of a mere 3 percentage points. But their suggestion that two wrongs make one right has little merit. Indeed, most of the wealth that is being taken out of the denominator should instead be added to the numerator. In the limit, if all durables belonged in life cycle wealth, the correction would come to an increase in life cycle wealth and corresponding decrease in the inherited share of 14 percentage points. Lacking a more solid base, row 2 reports the correction based on this alternative.

The third row of Table 2B shows the effect of eliminating from their estimate of inheritance the capitalized earnings from bequests. As was shown above, the exclusion of these earnings reduces the value of transfer wealth by 47 percent. Accordingly, in row 3, the share of inherited wealth is reduced to only $35\frac{1}{2}$ percent. As can be seen from row 4, this is still some 16 percent higher than my preferred estimate of the share of inheritance—around 20 percent—based on probate statistics.

The bulk of this difference is probably accounted for by Kotlikoff and Summers' inclusion of the imputed capitalized consumption of all dependent persons 18 years of age and over. There is, unfortunately, no way of estimating directly how much this inclusion adds to the conventional measure of inherited wealth, but several considerations suggest that the addition must be appreciable. To begin with, a large portion of those between 18 and 24 are still dependent, especially at the younger end of that spectrum. For instance, in 1970 about two-fifths of those aged 18–19 and one-third of those aged 18 to 24 were not in the labor force; the fraction of dependent persons would, presumably, be substantially higher. Another suggestive item is provided by the Kotlikoff and Summers estimate that expenditures for college education, which could hardly be a major component of all expenditure on dependents aged over 18, represents just over 20 percent of the total flow of bequests. In terms of Table 2B, this college adjustment alone would reduce the estimate in row 3 by 7.8 points. The large role of dependent consumption is also supported by two Figures (1 and 2) provided in K & S (1981) which show, for selected cohorts, the life path of income and consumption. According to these graphs, the cohort of 1910 is estimated to have life cycle dissaving for the first 50 years of its life, while the cohort of 1940 saved nothing over a similar span. Similar results are reported for other cohorts in the paper cited.

These results, which provide the foundations for the negligible accumulation of life cycle wealth, are inconsistent with information from many other sources using the household as the basic unit, and the conventional definition of saving. First, available information from many surveys indicates that households have, on average, substantial saving and net worth at least after age 25 (for example, Ando & Kennickell, 1985, Table II.1). Second, such saving is consistent with the fact that wealth rises fairly smoothly between age 25 and age 45 (according to Projector and Weiss, 1964, wealth rose in 1962 by roughly one thousand dollars per year of age; see also Ando & Kennickell, 1985). This rise in wealth cannot be attributed to inheritance to any significant extent, since, as one would expect, the receipt of important inheritance is rare before age 45 (Projector and Weiss, 1964, Table A32). For the same reason, the saving of these younger age groups cannot be reasonably attributed to the return on inherited wealth.

The residual of row 4 and the inconsistencies noted above may, of course, also reflect the entirely different data as well as methodology employed by Kotlikoff and Summers, for it is obvious that each method is affected by the many somewhat arbitrary auxiliary assumptions that need to be made. This problem particularly affects the Kotlikoff and Summers study, considering the large number of imputations and assumptions it involves, from age profiles, to return on capital, to the treatment of interspousal transfers. Their results are particularly sensitive to errors and assumptions affecting saving in the early years, for a difference in early saving affects wealth at every later age, and increasingly so, as the saving gets capitalized. This conjecture is supported by tests on the sensitivity of their results to variations in the auxiliary assumptions, which they report in their 1981 paper.

These considerations lend support to the claim that the bulk of the discrepancy between their estimate of 80 percent and the earlier consensus figure of around 20 percent is attributable to differences in the definitions, with the remainder accounted for by the unavoidable imprecision of all estimates. This inference receives further support from the estimates of Ando and Kennickell (1985) cited earlier. Using the same methodology of estimation as Kotlikoff and Summers but relying on the conventional definition of self-accumulated wealth as household actual income minus consumption, they arrive at a life cycle share much lower than Kotlikoff and Summers and broadly similar to the figures reported in Table 1.

Assessing the Merits and Shortcomings of Alternative Definitions. There are two basic definitional differences: how to define life cycle (hump) accumulation, and how to treat the return on inherited wealth. With respect to the first issue, the Kotlikoff and Summers redefinition was shown to have the effect of subtracting from the standard measure of hump wealth the capitalized value of all expenditure imputed to dependents age 18 and over (as well as all minor gifts and contributions in support of another household).

This redefinition has little merit, at least when the focus is on the effect of inheritance and gifts on the stock of (nonhuman) wealth. First, the supposed transfers (and contributions) are not in the nature of either bequests or major gifts. Second, they go to pay for current consumption and do not represent an addition to the assets of the recipient or society. Third, the downward adjustment implied by the Kotlikoff and Summers procedure depends on the choice of the critical age of independence, and as a result their share of life cycle wealth in total wealth can be made to rise or shrink by the fairly arbitrary choice of the critical age. Finally, these imputed transfers are quite different in nature from bequests and major gifts because, unlike these transfers, they would be hard to modify through policy actions, and, even if modifications were attainable, the effect on wealth could be very different.

One consideration that may lend some attractiveness to the Kotlikoff and Summers definition is that the expenditure on dependents that they treat as transfers includes the outlays for college education. One may feel that these large outlays are bequests of a sort, since they take the form of an investment in "human capital." But this consideration would be relevant mainly for other issues, such as the hereditary transmission of economic inequality or the contribution of transfers to total capital—nonhuman and human. But in this case the denominator of the ratio should include human capital, too. Furthermore, the numerator should include many other expenditures on human capital, not necessarily only on behalf of dependents 18 years old and over—like all private schooling. And why should the line be drawn at schooling and not include all expenditure on children? But clearly that would be an entirely different story. To answer the question this paper began by posing, I submit that no customary expenditure on dependents should be treated as a transfer.

The next question is whether inherited wealth should be defined to include the capitalized earnings thereon over the lifetime of the recipient. I have chosen to exclude it on two grounds. First, treating earnings on bequests as income conforms to the generally accepted definition of saving as income minus consumption. It also conforms to the usual definition of life accumulation as bequests left, minus bequests received (adjusted for inflation). By contrast, under Kotlikoff and Summers's definition, a person leaving as much as he received would, if the real rate of return were positive, be counted as a dissaver to an extent depending on the rate of return. The second reason is that one can measure directly what bequests have been received, but there is no direct way of telling whether some years later the wealth of the recipient will be larger by the capitalized value of the bequests, or whether instead the recipient will have consumed some or all of the return or even some of the principal.

Kotlikoff and Summers might object to this definition in that life cycle income as I am defining it includes the income from bequests, which makes life cycle saving depend on bequests received. The measure of life cycle saving they advocate (labor income minus consumption) is, instead, independent of bequests. In addition, this measure, since it requires subtracting the return on bequests from earnings and saving, provides a justification for adding capitalized returns to the stock of bequests received.

At first sight, this may seem like a persuasive argument; yet it suffers from a major flaw. While it is true that the measure I am supporting will generally vary with the size of bequests, theirs is also not independent of bequests received as long as these have effects on consumption (or labor supply). Their measure would be appropriate only in the limiting case in which consumption is absolutely unaffected by bequests received—that is, if all returns were saved. But consider the polar case in which all the returns are consumed; then it is the measure of bequests and hump saving I am advocating that is the appropriate one.

The Contribution of the Bequest Process to Total Wealth

What lesson can be drawn from the above considerations? They clearly suggest that the share of inherited wealth, whether capitalized or not, does not provide a valid answer to the question: *how much does the bequest motive contribute to society's total wealth?* Kotlikoff and Summers have correctly pointed out that this issue is the really interesting one. They have further proposed measuring the impact of bequests on total wealth by the elasticity of total wealth with respect to flow of bequests: that is, by the percentage change in total wealth resulting from a 1 percent change in the bequests' flow. This elasticity, which may be labelled the "true measure of importance of bequests," can be written as

$$\eta = \frac{\Delta W}{\Delta B}\frac{B}{W} = \frac{\Delta W}{\Delta T}\frac{T}{W}. \tag{3}$$

Here, the second equality follows from the consideration that the stock of inherited wealth T is proportional to the annual flow of bequests B (though the proportionality factor depends on the definition). The first equality, on the other hand, shows that (given the flow of bequests) the true measure of importance is independent of how one chooses to define the share, as must be the case since the elasticity is, in principle, an observable fact. The second equality brings to light a simple relation between the "true" measure of importance and the measures of the share of bequests with which we have been concerned so far. Specifically, let T^j stand for any stated measure of aggregate wealth, such as T^*, the definition used by Kotlikoff and Summers, or T, the definition I am advocating. Then the corresponding measure of the share, T^j/W, will be an upward biased measure of importance if $\Delta W/\Delta T^j$ is less than one; it will be downward biased in the opposite case.

No measure of the aggregate inherited wealth presented thus far is likely to measure the desired elasticity correctly for several reasons. In the first place, when the economy is growing, the share of wealth received by bequests must tend to underestimate the contribution of bequests to total wealth by not allowing for the effect of accumulation earmarked for future bequests. Some simulations carried out by Kennickell (1984) suggest that the shortfall of the inherited share as a measure of importance rises quite rapidly as growth increases. Second, there is reason to believe that $\Delta W/\Delta T$ is not unity, however it is defined, simply because beneficiaries of bequests will probably change their consumption by an amount positive, but less than the full return on bequests.

Measuring the Elasticity

Consider first the case where the stock of inherited wealth is T^*, defined in accordance with Kotlikoff and Summers. These authors have endeavored to estimate the value of $\Delta W/\Delta T^*$. Their calculations are based on some very special and rather arbitrary assumptions about preferences (namely that the utility is additive, separable in consumption and leisure and logarithmic) and on the even more questionable assumption that all transfers—be they inheritances or major gifts or family expenditure in support of members over 18—have identical effects on wealth. Nonetheless, the results should provide a general indication of magnitude.

They find that $\Delta W/\Delta T^*$ depends almost exclusively on the difference between the interest rate and the growth rate $(r - n)$, decreasing as that figure increases for reasons that will be discussed presently. For a value of $(r - n)$ consistent with their estimates of r and n, namely .01, they find that $\Delta W/\Delta T^* = .7$. This result, together with equation (3), means that the share of bequeathed wealth, measured according to Kotlikoff and Summers definition, T^*/W, greatly *overstates* η, the true effect of bequests on total wealth, to wit, by 1/.7, or over 40 percent.

I have endeavored also to estimate $\Delta W/\Delta T$, but by an alternative procedure which is less dependent on specific assumptions about preferences, and was initially inspired by Darby (1979). He proposed identifying the "true" life cycle component of wealth as the amount of wealth that would be in existence in society if households accumulated just enough assets to enable them to finance their observed (average) retirement consumption --- with accumulation up to retirement, and decumulation thereafter, occurring at a constant rate.

Darby has applied this method to U.S. data around 1966. The rate of consumption to be financed during retirement was estimated using data from the 1967 Survey of Economic Opportunities and the portion of this to be financed through "life cycle accumulation" was obtained by subtracting, from consumption other sources of income, such as labor income and Social Security. The retirement fund was assumed to earn the rate of return of 4.5 percent which was adopted in our previous calculations. This method estimates the share of "life cycle wealth" at only 23 percent, implying that over three-fourths of wealth is bequest-related.

But Darby's approach, for all its ingenuity, cannot provide much useful information because of the entirely arbitrary nature of the underlying assumptions. His assumption that accumulation for retirement occurs smoothly while accumulation for bequests is the jagged residual (see Darby's figure 12, p. 37) is a caricature of consumption smoothing based on life cycle utility maximization, which implicitly treats most hump wealth as bequest-related wealth.

This consideration suggests modifying his approach by replacing the ad hoc assumption that consumers smooth the accumulation of retirement provisions with the "rational" assumption of consumption smoothing. Specifically, given the amount to be bequeathed, how would that amount be accumulated by a person choosing an optimal life consumption path, subject to the constraint imposed by available lifetime resources? For the representative household, these resources consist of lifetime earnings plus bequests received, less the amount the household intends to bequeath. Steady state considerations imply that, on the average, bequests left will exceed those received by the growth factor e^{ng}, where n is the rate of growth and g the age gap between donor and recipient. Wealth holding due to bequests can then be computed as the difference between the path of wealth with and without bequests.

As long as the optimal consumption follows a smooth path (a constant rate of growth) which is consistent with commonly assumed additive utility functions, one can readily show that accumulated wealth due to bequests will rise smoothly to an amount equal to the difference between the bequests left at death and those received, capitalized from the date of receipts.

If the preferred consumption path grows at the rate c, the annual increment to wealth due to bequests in the year τ can be shown to be:

$$\Delta s_\tau = \frac{Ae^{c\tau}(r-c)}{1-e^{(r-c)L}} \tag{4}$$

where L is length of life and A is the present value, at the beginning of life, of the difference between the bequest received and left. The path of wealth is the cumulant of Δs_τ plus the amount of the bequests once received (capitalized).

Equation 4 describes the path for a single household. To obtain the aggregate amount of bequests-related wealth at a given point of time, we must sum over the wealth of each cohort present, allowing for the fact that, because of growth, the cohorts of age τ can expect to receive and leave bequests which are larger than those left by the one currently deceased by the factor $e^{(L-\tau)n}$. National wealth is the summation of wealth over the cohorts adjusted for mortality. Finally, taking into account that the steady state bequests must grow at the rate n, bequests received will be Be^{-ng}. This formulation permits us to derive an expression for aggregate bequests-related wealth, ΔW, in terms of the current flow of bequests left, B, and the parameters c, n, r, g, p_t and L, where p_t is the force of mortality at age t. Taking as an illustration the case in which mortality is zero until age L and 1 at L, the result can be written as

$$\Delta W = \Delta T^{*}(1-D) \tag{5}$$

where T^* is the stock of bequeathed wealth according to K&S's definition as given by equation (2), and D takes the somewhat lengthy expression:

$$D = \frac{e^{-(r-n)L}}{1 - e^{-(r-c)L}}\left[e^{(r-n)L} - 1 - \left(1 - e^{-(n-c)L}\right)\frac{r-n}{n-c}\right]. \tag{6}$$

Equations (5) and (6) have a number of plausible and interesting implications. Consider first the case where the economy is stationary: r is zero and so is n. In this case, since bequests are passed on unchanged, they have no effect on income, consumption, or life cycle earnings. Thus, $T^* = T$, and $D = 0$, and by (5), $\Delta W/\Delta T^* = \Delta W/\Delta T = 1$. In this case, the two definitions of the share of inherited wealth coincide and correctly measure "importance."

Consider next the case of a stationary economy, but in which $r > 0$. Here the receipt of inheritance has, on average, a favorable income effect, as it earns interest but requires no additional accumulation; one can show that, for this reason, my definition of the share, T/W, has an upward bias in measuring η, but using Kotlikoff and Summers' measure, the upward bias is clearly even greater. On the other hand, a positive growth tends, through the "accumulation" effect, to impart a downward bias to either measure of the share. Thus, r and n work in opposite and, in fact, offsetting ways, as is apparent from the fact that in equation (6), r and n often appear in the form $(r - n)$. So, the Kotlikoff and Summers measure of the share is upward biased as long as $r - n > 0$, and their parameter choices imply that $r - n$ is 0.01. This is sufficient to impart a strong upward bias to T^*/W as a measure of η. Indeed, assuming again for g and L the values of 25 and 55 respectively, one finds from (6) that $D = 0.3$, and hence, from (5), $\Delta W/\Delta T^* = 1 - D = 0.7$. (Note that this corresponds with the estimate of dW/dT^* reported by Kotlikoff and Summers.) Then, from (3)

$$\eta = .7\frac{T^*}{W} \tag{7a}$$

confirming that the Kotlikoff and Summers share overestimates the true measure of importance by over 40 percent. On the other hand, for the alternative measure of the importance of inherited wealth that I am advocating, equation (5) implies

$$\frac{\Delta W}{\Delta T} = \frac{\Delta W}{\Delta T^*}\frac{\Delta T^*}{\Delta T} = \frac{\Delta W}{\Delta T^*}\frac{T^*}{T} = (.7)(1.7) = 1.19$$

where T is computed from equation (1) and T^* from (2). Hence:

$$\eta = 1.19\frac{T}{W}. \tag{7b}$$

Thus, as expected our share underestimates "importance" (because it neglects the accumulation effect due to growth), but the bias is fairly small (in the relevant range of parameters).[6]

[6] The calculations reported above assume a zero value for the rate of growth of consumption c. But the value of $\Delta W/\Delta T^*$ is not significantly affected by variations in c in the relevant range. Thus, for $c = .02$, $\Delta W/\Delta T^* = .723$, and $\Delta W/\Delta T = .723 \times 1.7 = 1.23$.

Estimates of the Elasticity. As for the actual value of the elasticity η, it can be computed either starting from our definition of the share T/W and using (7b) or from Kotlikoff and Summers T^*/W and using (7a)—and the result should coincide, at least as long as the shares are based on the same definition of bequests. As shown in Table 2A, my estimate of T/W is 17 percent (line 4) implying $\eta = .17 \times 1.19 = 20$ percent. The Kotlikoff and Summers estimate of T^*/W is .46 percent (Table 2A, line 1) which, however, falls by 41 percent if one accepts the hypothesis that bequests are typically left at death. Using (7a), this would imply an elasticity of .29. However, the estimates are still comparable because Kotlikoff and Summers are including in their flow of bequests and gifts the imputed consumption of dependents over 18 years of age.

As argued earlier, this expenditure should not be included in the bequest flow. With respect to the present problem, there is one further reason for exclusion, namely that no allowance was made for such flows in our (or as far as I can see, in Kotlikoff and Summers's) calculation of $\Delta W/\Delta T$.

If we accordingly eliminate from T^*/W the component of line 3, which reflects this expenditure, the share falls to 32 percent, implying η around .22, pretty close to our .2. But both estimates may tend to underestimate η because the Kotlikoff and Summers measure of the flow of bequests, which ignores nonintergenerational transfers, is presumably too narrow. Probate data and the results of Ando and Kennickell (1985) suggest a larger value for T/W, say between 1/5 and 1/4. Using (7b), this would imply an elasticity of between .25 and .30.

There remains to consider Kotlikoff and Summers's alternative estimate of T^*/W, based on the capitalization of life cycle saving, which is analyzed in Table 2B. If we correct as well as we can for the treatment of durable goods, as described earlier, line 2 shows that T^*/W would come to .67, which, together with (7a), implies $\eta = .47$. But, the above value of T^*/W is greatly biased upward by the inclusion of the consumption of family members over 18. Unfortunately, there is no way to measure that bias at present. Correcting for only a portion of the above expenditure—namely expenditure on college education—would lower T^*/W from .67 to .57 (cf. Table 2A) and bring η down to .4. One can conjecture that the full correction would bring η pretty close to the upper range of .3.

We can, therefore, conclude that when we focus on the "importance" of bequests as measured by the elasticity, all results point to a value of η of up to 30 percent, give or take a few points.

The Precautionary Motive and the Importance of Bequested Motivated Wealth

However, this value of .3 overestimates the contribution to wealth of inheritances and gifts related to the bequest motive because a substantial portion of the observed bequest flow undoubtedly reflects the precautionary motive arising from the uncertainty of the time of death. Indeed, in view of the institutional obstacles of dying with negative net worth, people tend to die with some wealth, unless they can manage to put all their retirement reserves into life annuities. In the absence of annuities, the wealth left behind will reflect risk aversion and the cost of running out of wealth (besides the possible utility of bequests).

This point has been elaborated in particular by Davies (1981) (see also Hubbard, 1984) who has shown that, for plausible parameters of the utility function including a low intertemporal elasticity of substitution, the extent to which uncertainty of life depresses the propensity to consume increases with age. As a result "uncertain life time could provide the major element in a complete explanation of the slow decumulation of the retired," relative to what would be implied by a standard Life Cycle Hypothesis model.

Clearly, bequests originating from the precautionary motive are quite different by nature from those dictated by the bequest motive. Indeed, they belong with pure life cycle accumulation since they are determined by the utility of consumption, and furthermore, the surviving wealth must tend, on the average, to be proportional to life resources.

However, using the precautionary motive as an explanation for life cycle wealth does run into some problems. If the purpose of the wealth accumulated at retirement was really to support consumption, then given uncertainty of life, risk aversion and the availability of annuity contracts, why don't more households use most of their wealth to buy annuities?

This criticism is important, but a number of counterarguments have been proposed. First, a fair amount of consumer wealth at retirement is, in fact, in the form of annuities: namely, all that is accumulated in the form of claims on pension funds and Social Security. Indeed, one might turn the above question around and ask: if households accumulate primarily to leave bequests, why have pensions and Social Security met with so much success and growth? Second, as has been pointed out in particular by Friedman and Warshawsky (1985a, 1985b), one important factor discouraging the purchases of annuities is the very unfavorable rates which are currently offered on such contracts (estimated by these authors at 4 to 6 percent below market rates of return). Of this "load," only one portion (around 1.5 percent) can be attributed to adverse selection (the fact that those who chose to buy annuities tend to have above average life).

Friedman and Warshawsky have shown that this unfavorable load factor is probably not large enough to lead a person to choose self-insurance if that person derived no utility whatever from bequest. However, the actual situation is different since, as a rule, households do derive some utility also from bequests (Masson, 1986). Friedman and Warshawsky have shown that, in this case, under plausible assumptions, the extra load factor can account for the rarity of private annuities. They also report that under the same conditions, the household would tend to leave an amount of bequest relative to terminal consumption that appears broadly consistent with the observed behavior of wealth as a function of age, indictating that this behavior might be accounted for even in the absence of a pure bequest motive.

The Importance of the Pure Bequest Motive

It would be interesting to obtain some estimate of the importance of purely bequest-motivated transfers. A certain amount of evidence suggests that the pure bequest motive—the accumulation of wealth entirely for the purpose of being distributed to heirs and not be used for own consumption—affects a rather small number of households, mostly located in the highest income and wealth brackets.

First, the recent study of Hurd (1986) supports the hypothesis that the bequest motive is not important for a broad cross section of households. He starts from the reasonable hypothesis that if the true bequest motive is an important source of terminal wealth, then retired households with living children should have more wealth and should save more (dissave less) than childless ones. It is found that, in fact, those with children have less wealth and, by and large, dissave the same fraction of (marketable) wealth. The first result is fully consistent with the standard no-bequest life cycle consumption smoothing because for given life resources the "cost" of raising children reduces the retired consumption of parents. But the second result is indeed hard to reconcile with a significant pure bequest motive. It is, however, consistent with the finding of Projector and Weiss (1964) that only 3 percent indicated they were saving "to provide an estate for the family." At the same time, the proportion rose with wealth, reaching one-third for the top class (over half a million 1963 dollars). Thus, the bequest motive seems to be concentrated in the highest economic classes.

This hypothesis is supported by the finding of Menchik and David that for (and only for) the top 20 percent of the distribution of estimated life resources bequests rise proportionately faster than total resources.

This result suggests that the share of bequests due to the pure bequest motive is likely to be well below one, even allowing for the fact that the wealth of those with pure bequest motives may be a sizable part of the total. Recalling that our estimate of elasticity of wealth with respect to the entire flow of bequests came to somewhere around .3, it would seem safe to conclude that the importance of pure bequest motivated transfers, as measured by the elasticity of wealth with respect to that flow, is very unlikely to exceed one-fifth or so.

Summary and Conclusions

Clearly, part of the private wealth held at any time reflects hump or life cycle wealth and part reflects wealth transmitted through inheritances and major gifts. The interesting question is: how large is each component? The available evidence, reported in seven studies largely relying on independent methods but using broadly similar, customary definitions, consistently indicates that the share of wealth received by transfer does not exceed one-fourth. One recent contribution, that of Kotlikoff and Summers (1981), based on different definitions and partly on a different methodology, has arrived at a much larger share of 45 to 80 percent.

It has been shown that the differences between these and the other estimates reflect mainly definitional differences which have the effect of substantially increasing K&S's measure of the share. These differences consist in adding to the bequests and major gifts received: 1) the capitalized value of the earning on the inheritance since the time of receipt, and 2) the capitalized value of the expenditure on family members over 18 years in age.

But these definitional differences lose relevance when we focus on the elasticity of wealth with respect to the flow of bequests. It is first argued that, with respect to this issue, Kotlikoff and Summers's treatment of expenditures on family members over 18 years of age as bequests is not appropriate. Once that component is eliminated, the two measures of the share are shown to give rise to roughly similar values of the elasticity; a 10 percent decline in the flow of bequests might result in a decline in wealth of the order of 3 percent, more or less.

Even this figure overestimates the role of bequest motivated transfers, which seem to play an important role only in the very highest income and wealth brackets. Some portion of bequests, especially in lower income brackets, is not due to a pure bequest motive but rather to a precautionary motive reflecting uncertainty about the length of life, although it is not possible at present to pinpoint the size of this component.

References

Ando, Albert, "The Savings of Japanese Households: A Micro Study Based on Data from the National Survey of Family Income and Expenditure, 1974 and 1979," Economic Planning Agency, Government of Japan, 1985.

Ando, Albert, and A. Kennickell, "How Much (or Little) Life Cycle is There in Micro Data? Cases of U.S. and Japan," paper presented at Istituto Bancario San Paolo di Torino Conference to Honor Franco Modigliani, in Martha's Vineyard, Massachusetts, September 19–21, 1985. Conference proceedings forthcoming.

Avery, R. B., G. E. Elliehausen, G. B. Canner, and T. A. Gustafson, "Survey of Consumer Finances, 1983: A Second Report," *Federal Reserve Bulletin*, December 1984, *70*, 857–68.

Barlow, Robin, H. E. Brazer, and J. N. Morgan, *Economic Behavior of the Affluent*. Washington, D.C.: The Brookings Institution, 1966.

Bernheim, B. Douglas, "Dissaving after Retirement: Testing the Pure Life Cycle Hypothesis," NBER Working Paper #1409, July 1984.

Board of Governors of the Federal Reserve System Flow of Funds Section, Division of Quarterly Statistics, *Balance Sheets of the U.S. Economy 1945–1980*, October 1981.

Darby, Michael R., *The Effects of Social Security on Income and the Capital Stock*. Washington, D.C. The American Enterprise Institute, 1979.

David, Martin, and P. Menchik, "Distribution of Estate and Its Relationships to Intergenerational Transfers," In Department of the Treasury, Internal Revenue Service, Statistics of Income Division, *Statistics of Income and Related Administration Record Research: 1982*, October 1982.

Davies, J., "Uncertain Life Time, Consumption and Dissaving in Retirement," *Journal of Political Economy*, June 1981, *89*, 561–77.

Diamond, Peter A., and J. A. Hausman, "Individual Retirement and Savings Behavior," *Journal of Public Economics*, February–March 1985, *23*, 81–114.

Evans, O. J., *The Life Cycle Inheritance: Theoretical and Empirical Essays on the Life Cycle Hypothesis of Saving*. Unpublished Ph.D. dissertation, University of Pennsylvania, 1981.

Evans, O. J., "Empirical Tests of the Life Cycle Hypothesis: Comment," *American Economic Review*, March 1984, *74*, 254–257.

Federal Reserve Board, Flow of Funds Section, Division of Research and Statistics, *Balance Sheets for the U.S. Economy, 1945–80*, 1984.

Feldstein, Martin, "Social Security and Private Savings: International Evidence in an Extended Life-Cycle Model." In Feldstein, M., and R. Inman, eds., *The Economics of Public Services*. London: Macmillan, 1977.

Fisher, Irving, *The Theory of Interest*. New York: Macmillan, 1930.

Fisher, J., "The Economics of an Aging Population, A Study in Income, Spending and Savings Patterns of Consumer Units in Different Age

Groups, 1935-36, 1945-1949." Unpublished dissertation, Columbia University, New York, 1950.

Friedman, Benjamin M., and M. Warshawsky, "The Cost of Annuities: Implications for Saving Behavior and Bequests," National Bureau of Economic Research Working Paper #1682, August 1985a.

Friedman, Benjamin M., and M. Warshawsky, "Annuity Prices and Saving Behavior in the United States," National Bureau of Economic Research Working Paper #1683, June 1985b.

Harrod, R. F., *Towards a Dynamic Economics.* London: Macmillan, 1948.

Hubbard, Glenn R., "'Precautionary' Saving Revisited: Social Security, Individual Welfare, and the Capital Stock," National Bureau of Economic Research Working Paper #1430, August 1984.

Hubbard, Glenn R., "Pension Wealth and Individual Saving: Some New Evidence," *Journal of Money, Credit and Banking,* May 1986, *XVIII(2),* 167-78.

Hurd, Michael D., "Savings and Bequests," National Bureau of Economic Research Working Paper #1826, January 1986.

Keynes, J. M., *General Theory of Employment, Interest and Money.* New York: Harcourt, Brace, 1936.

Kennickell, A., "An Investigation of Life Cycle Savings Behavior in the United States," unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, 1984.

King, M. A., and L-D. L. Dicks-Mireaux, "Asset Holdings and the Life-Cycle," *Economic Journal,* June 1982, *92,* 247-67.

Kotlikoff, L. T., and L. H. Summers, "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy,* August 1981, *89,* 706-32.

Kotlikoff, L. T., and L. H. Summers, "The Contribution of Intergenerational Transfers to Total Wealth: A Reply," National Bureau of Economic Research Working Paper #1827, 1986.

Long, Clarence D., "The Labor Force Under Changing Income and Employment," National Bureau of Economic Research, General Series 65, Princeton University Press, 1958.

Lydall, Harold, "The Life Cycle in Income, Saving and Asset Ownership," *Econometrica,* April 1955, *23,* 131-50.

Masson, Andre, "Cohort Analysis of Wealth-Age Profiles Generated by a Simulation Model in France (1949-1975)," *Economic Journal,* March 1986, *96.*

Menchik, Paul, and M. David, "Income Distribution, Life Time Saving and Bequests," *American Economic Review,* September 1983, *73,* 672-90.

Mirer, T. W., "The Wealth-Age Relationship among the Aged," *American Economic Review,* June 1979, *69,* 435-43.

Modigliani, Franco, "The Life Cycle Hypothesis of Saving, Twenty Years Later." In Parkin, M., ed., *Contemporary Issues in Economics.* Manchester: Manchester University Press, 1975, pp. 2-36.

Modigliani, Franco, "Measuring the Contribution of Intergenerational Transfers to Total Wealth: Conceptual Issues and Empirical Findings," Paper presented at a seminar on Modeling the Accumulation and Distribution of Personal Wealth, in Paris, France, September 1984.

Modigliani, Franco, and Albert Ando, "Tests of the Life Cycle Hypothesis of Savings: Comments and Suggestions," *Bulletin of the Oxford University Institute of Statistics,* 1957, 99-124.

Modigliani, Franco, and Richard Brumberg, "Utility Analysis and Aggregate Consumption Functions: An Attempt at Integration." In Abel, A., ed., *The Collected Papers of Franco Modigliani.* Cambridge, MA: The MIT Press, 1980.

Modigliani, Franco, and Richard Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data." In Kurihara, K. K., ed., *Post-Keynesian Economics,* New Brunswick: Rutgers University Press, 1954, pp. 388-436.

Modigliani, Franco, and Arlie Sterling, "Determinants of Private Saving with Special Reference to the Role of Social Security Cross-Country Tests," In Modigliani, F., and R. Hemming, eds., *The Determinants of National Saving and Wealth.* New York: St. Martins Press, 1983, pp. 24-55.

Morgan, J. N., M. H. David, W. J. Cohen, H. E. Brazer, *Income and Welfare in the United States.* New York: McGraw-Hill Book Company, Inc., 1962.

Projector, Dorothy, *Survey of Changes in Family Finances.* Washington, D.C.: The Board of Governors of the Federal Reserve System, 1968.

Projector, Dorothy, and Weiss, Gertrude S., *Survey of Financial Characteristics of Consumers.* Washington, D.C.: The Board of Governors of the Federal Reserve System, 1964.

Royal Commission on the Distribution of Income and Wealth, Report No. 5, Third Report on the Standing Reference.

Shorrocks, A. F., "The Age-Wealth Relationship: A Cross-Section and Cohort Analysis," *Review of Economics and Statistics,* May 1975, *57,* 155-63.

White, Betsy Buttrill, "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis and Aggregate Household Saving," *American Economic Review,* September 1978, *68,* 547-60.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 41–58

Intergenerational Transfers and Savings

Laurence J. Kotlikoff

In recent years the role of intergenerational transfers in the process of wealth accumulation has been the subject of substantial empirical and theoretical analysis. The key question stimulating this research is: What is the main explanation for savings? Is it primarily accumulation for retirement as claimed by Albert Ando, Richard Brumberg, and Franco Modigliani in their celebrated Life Cycle Model of Savings? Is it primarily intentional accumulation for intergenerational transfers? Or is it primarily precautionary savings, much of which may be bequeathed because of imperfections in annuity markets?

The answer to the savings puzzle has many policy implications; certain tax structures are much more conducive to some types of savings than others, and certain government insurance programs might appear less attractive if precautionary motives are the main explanation of savings. Knowledge of the primary savings mechanism would also provide the key to understanding the distribution of wealth.

Solving the savings puzzle requires first collecting the pieces and then seeing how they fit together. A major piece of the puzzle is understanding the quantitative importance of intergenerational transfers to the accumulation of wealth. As I argue below, there is strong evidence that intergenerational transfers play a very important and perhaps dominant role in U.S. wealth accumulation. This does not mean, however, that intentional saving for gifts and bequests is the main motive for savings. Significant intergenerational transfers could also arise in the Life Cycle Model in the absence of well-functioning private annuity markets or close substitutes for such markets. In such a setting bequests would be involuntary and potentially quite sizeable (Bevan and Stiglitz, 1979; Kotlikoff and Spivak, 1981; Davis, 1981;

■ *Laurence J. Kotlikoff is Professor of Economics, Boston University, Boston, Massachusetts, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts.*

Kotlikoff, Shoven, and Spivak, 1986, 1987; Abel, 1985, Eckstein, Eichenbaum, and Peled, 1983). Let us first look at the evidence on the importance of intergenerational transfers and then turn to the deeper question of why such transfers arise.

The Importance of Intergenerational Transfers to Savings

There are six types of evidence concerning the importance of intergenerational transfers to savings. First comparisons of total U.S. wealth with life cycle wealth, defined as the amount of U.S. wealth there would be in the absence of any net intergenerational transfers. The difference between total wealth and life cycle wealth is defined as transfer wealth. These calculations use age-earnings and age-consumption profiles as well as other data, but they do not use data on transfers since they are concerned with what wealth would be in the absence of transfers. Second, the calculation of transfer wealth as defined above using steady state assumptions and the limited reported data on the flow of transfers. Third, zero transfer, life cycle simulation models that attempt to reproduce in a realistic manner actual wealth to income ratios or actual wealth distributions. Fourth, analysis of the rate of asset decumulation of the elderly. Fifth, evidence from annuity markets. Sixth, historical evidence concerning the correlation of saving rates and changes in the length of retirement.

Each of these types of evidence suggests an important role for intergenerational transfers in savings.

The Calculation of Life Cycle and Transfer Wealth Components

Kotlikoff and Summers (1981) considered the following question: Are the U.S. data broadly consistent with the view that intergenerational transfers play a negligible role in U.S. wealth accumulation? Stated differently, can one reject the null hypothesis that the Life Cycle Model fully explains U.S. wealth? To address this question we divided total wealth W into two components, life cycle wealth L and transfer wealth T and defined life cycle wealth of a cohort as the sum over each age of the accumulated difference between past streams of labor earnings and consumption. Total life cycle wealth equals the sum over cohorts of each cohort's life cycle wealth. Let E and C stand, respectively, for the sum over cohorts of all past accumulated earnings and all past accumulated consumption, then

$$W = L + T, \quad \text{and} \quad L = E - C. \tag{1}$$

With this definition of life cycle wealth L, transfer wealth T equals the sum over cohorts of the sum of accumulated net transfers received at each age. Calculating life cycle wealth requires knowledge of longitudinal age earnings and age consumption profiles for each cohort. These profiles were derived using data for the period 1900 through 1974 on cross section relative age earnings and age consumption profiles as well as aggregate earnings and consumption. The procedure involved distributing total earnings and consumption in each year by age and sex; combining these cross section age-sex matrices of earnings and consumption yields longitudinal profiles that are used to form E and C, past accumulated earnings and past accumulated consumption. In our initial paper Summers and I reported a value of life cycle wealth for 1974 of $733 billion, which is only 18.9 percent of total 1974 household wealth of $3,884 billion. We subsequently identified an error in the treatment of durables. In our reply to Modigliani (1984)—Kotlikoff and Summers (1986)—we indicate that a proper correction for durables raises the share of life cycle wealth to 21.9 percent.[1]

While the 21.9 percent is quite small, it is probably an overestimate of the life cycle share of total wealth. In order to generate at least some positive value for life cycle wealth, Summers and I adjusted upwards the standard estimates of the labor income of the self-employed by 20 percent. Since life cycle wealth is the accumulated difference between labor earnings and consumption, larger labor earnings of the self-employed implies a larger calculated value of life cycle wealth. The ratio of self-employed workers to employees was substantially larger in the prewar period than it is today. Hence, the calculated value of life cycle wealth is fairly sensitive to the estimation of the labor earnings of the self-employed. Using standard estimates would reduce life cycle wealth by about $700 billion. In addition, correcting several other intentional biases in our calculation would produce negative values for life cycle wealth. These include our assumption of a quite high ratio of female to male earnings, our assumption of zero earnings after age 75, and our assumption that the age-consumption profile is flat, rather than declining, after age 75.

It may be useful to repeat our basic explanation for why life cycle wealth is so small in the United States. Unlike simple classroom depictions of hump saving in which the age consumption is flat and the earnings profile rises to retirement, actual age earnings and age consumption profiles, such as those in Figures 1 and 2 which are reproduced from our paper, have essentially identical shapes and levels prior to at least age 45. Between ages 45 and 60 there clearly is some hump saving in that earnings profiles exceed consumption profiles; also after age 60 the age consumption profile clearly exceeds the age earnings profile. However, this pattern of hump saving and dissaving (at least relative to earnings) occurs quite late in the life cycle. Hence, one would not expect a large accumulation of life cycle wealth in the aggregate since the life cycle wealth of the more numerous generations below age 45 is so small. The simple fact is that consumption does not rise more rapidly through life than labor income.

An earlier study by Darby (1979) that influenced my study with Summers used cross section data on wealth earnings and consumption to divide current wealth holdings into a fraction that would be consumed and a fraction that would be transferred to succeeding generations. Darby inferred longitudinal age-consumption and age-earnings profiles from the cross section profile and concluded that at most 29

---

[1] The 26 percentage point durable correction referred to by Modigliani in this journal is based on highly suspect steady state assumptions that we do not regard as reasonable.

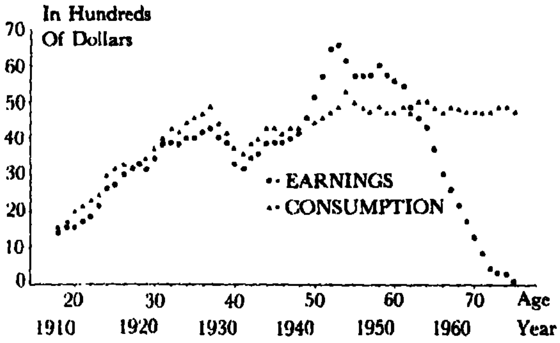


Fig. 1. Sum of male and female longitudinal average earnings and average consumption profiles, age 18 in 1910–age 82 in 1974

percent of U.S. private net worth is devoted to future consumption, with the rest destined for intergenerational transfer. White (1978) used aggregate data on the age structure of the population, age earnings and age consumption profiles along with a variety of parametric assumptions and concludes that the life cycle model can account for only about a quarter of aggregate saving. Though their accounting frameworks are somewhat different and though they use different data, and only cross section data at that, Darby and White reach essentially the same conclusion as Kotlikoff and Summers because the basic shapes of U.S. cross section age earnings and age consumption profiles and the longitudinal profiles that can reasonably be inferred from the cross section profiles are quite different from those of the textbook life cycle model.

Calculations of Life Cycle and Transfer Wealth Using Flow Data

The analyses just described directly calculate life cycle wealth and indirectly infer the stock of transfer wealth. Obviously it would be very useful to corroborate these results with direct evidence on intergenerational transfers. Kotlikoff and Summers

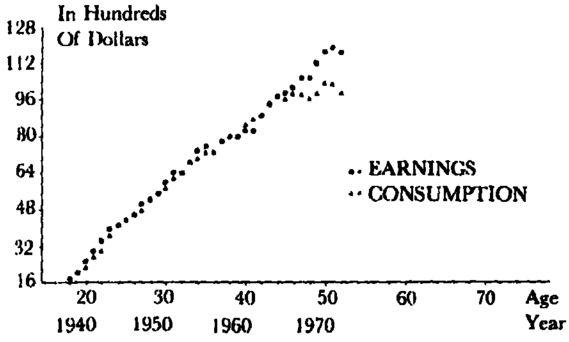


Fig. 2. Sum of male and female longitudinal average earnings and average consumption profiles, age 18 in 1940–age 52 in 1974. Reproduced by permission of the University of Chicago Press.

(1981) also examine the available flow data on intergenerational transfers. We presented a formula, which is valid only in the steady state, that relates the flow of transfers to the stock of transfer wealth.

There are three major problems with this method for calculating T. First, the available flow data on transfers clearly provide a lower bound estimate for total transfer flows; there are no data sources that systematically report intergenerational transfers made in the form of implicit and explicit gifts. Explicit gifts, which may be in kind as well as in cash, are clearly acknowledged as such by donors and recipients. Implicit gifts, such as making one's son an equal partner in a lucrative family business or providing low interest loans to children, may not be viewed as a gift by donors or recipients and would be hard to identify in a survey. Since the U.S. distribution of wealth is highly skewed, implicit gifts, while perhaps small in number, could be very large in value. Hence, any flow estimates of transfer wealth (particularly those of Modigliani in this journal, since he ignores all non-bequest intergenerational transfers), should be viewed as potentially seriously downward biased.

The second problem with using flow data to calculate T is that the assumption of a steady state may be far from justified. It may be, for example, that the flow of intergenerational transfers in relation to the scale of the economy was much greater in the 1920s than in the 1960s and 1970s. Not only must one assume that aggregate variables have been in steady state since at least the turn of the century, but one must also assume that the age distribution of transfers is time invariant.

The third problem is that the simple formula given in equation (2) relating transfer wealth T to the transfer flow (the formula incorporates Modigliani's (1984) correction for a small error) assumes that everyone dies at the same age D, that all transfers are received at the same age I, and that all transfers are made at the same age G. This is obviously unrealistic, and it is not clear what choice of these three ages best approximates reality. The appropriate choice of these ages depends on one's assumption about the steady state interest and growth rate of the economy, r and n, respectively. If half of transfers are received at age 20 and half at age 60, using 40 for approximate I would be inappropriate; transfers received at age 20 should receive more weight in the approximation formula because they are accumulated for a much longer period than transfers received at age 60 and because the accumulation formula is a nonlinear function of age.

$$T = \frac{t}{(r - m)} e^{(r-n)D} \left[1 - e^{(n-r)(G-I)}\right] e^{(n-r)I} \tag{2}$$

To illustrate the implication of the formula, Summers and I discussed an example in which D equals 55 (a real world age of death of 73 if the age of adulthood is 18), the age gap $(G - I)$ between those leaving bequests and those receiving them equals 30, and $(r - n)$ equals .01. Because of our algebraic error we did not assume a value for the age I that transfers are received. Using a value of I equal to 15 (a real world age of 33) seems justified when r exceeds n, because transfer wealth depends on the period of accumulation, and the appropriate approximation to I should be smaller if

r exceeds n. The resulting value of G of 45 also seems reasonable given the fact that not all transfers are made at death. A similar statement holds for the choice of the age gap $G - I$; thus it seems likely that the choice of 30 for $(G - I)$ is too small given that r exceeds n. But keeping $(G - I)$ at 30 and choosing I equal to 15, the factor multiplying the flow of transfers t in equation (2) is 39. Since the estimated 1974 flow of reported transfers is $45.5 billion, the "transfer flow" lower bound estimate for T is $1,771 billion, or 46 percent of 1974 household wealth. Note that using a potentially more appropriate age gap of $(G - I)$ of 45 would yield a lower bound value of T of $2455, or 63 percent of total wealth.

Modigliani's Critique and Calculations

In his paper in this journal, Modigliani focuses to a very large extent on two issues. The first is "flow" estimates of the importance of intergenerational transfers to savings, and the second is the proper definition of life cycle versus transfer wealth. Modigliani devotes little space to the main contribution of Kotlikoff and Summers (1981), the direct calculation of life cycle wealth. We devoted most of our paper to the direct calculation of life cycle wealth because, as we stressed, the "flow" approach overestimates life cycle wealth due to the absence of data on a variety of transfer flows. In addition, as mentioned, the flow approach requires invoking steady state and other simplifying assumptions that may not be valid.

Defining Life Cycle Wealth. Our definition of life cycle wealth is motivated by the following question: Are the U.S. data on labor earnings, rates of return, consumption, and wealth broadly consistent with the view that intergenerational transfers play a negligible role in U.S. wealth accumulation? Stated differently, can one reject the null hypothesis that the life cycle model without intergenerational transfers fully explains U.S. wealth? We defined life cycle wealth according to the theoretical prediction of the zero intergenerational transfer, life cycle model, namely as the sum over cohorts of the accumulated difference between past streams of labor earnings and consumption. We defined the difference between actual U.S. wealth and life cycle wealth as transfer wealth.

Modigliani focuses on a different definition of the two components of life cycle wealth. He also suggests that we have "redefined" the two components. While Modigliani asserts that our definition of life cycle wealth is non-standard, it is the definition used in the two previous extensive analyses by Tobin (1967) and Darby (1979) of the role of the pure life cycle model in U.S. wealth accumulation. This definition is also the one chosen by Hayashi (1986) in his inquiry into life cycle and intergenerational transfer saving in Japan. Indeed, the practice of comparing the wealth predicted by the zero intergenerational transfer life cycle model with total wealth dates to Modigliani and Brumberg (1954).

Modigliani prefers to define life cycle wealth as L_m, where L_m equals the sum over cohorts of the sum of saving at each age, with saving defined as income less consumption. The problem with his definition is that income may include capital income earned on previously received intergenerational transfers. Hence, since income itself may reflect intergenerational transfers, the sum of saving out of income cannot be used to test with maximum power the null hypothesis that the zero transfer life cycle model accounts for essentially all of U.S. wealth. Stated differently, under the null hypothesis $L_m = L$, but if there are significant intergenerational transfers L_m could be close to total wealth W, although both L_m and W would be substantially different from L, which is our definition of life cycle wealth. Indeed, Ando (1986) reports calculations of L_m equal to 65 percent of the 75 percent of W for the period 1960 to 1970, while our 1974 calculation of life cycle wealth is only 21.9 percent of 1974 wealth.

Once one finds the data are highly inconsistent with the zero transfer, life cycle formulation, a natural behavioral question to raise is: What would be the impact on U.S. wealth of eliminating all intergenerational transfers? We raised this economic, as opposed to accounting, issue on the first page of our 1981 paper, indicating how our definition and estimate of life cycle wealth could be used to address this unrealistic, but nonetheless interesting, counterfactual. The answer to this economic question is, of course, independent of accounting convention. It takes into account the changes in life cycle wealth that might arise if intergenerational transfers were eliminated. More precisely, in a world of significant transfers some fraction of consumption and leisure may be financed out of such transfers, and their elimination could affect the value of life cycle wealth. Our answer to the economic question, to which we still subscribe, was that totally eliminating intergenerational transfers would, in partial equilibrium, reduce U.S. wealth by at least 50 percent. This economic (as opposed to accounting) statement suggests a much more important role for intergenerational transfers than has generally been thought to be the case.

Bequests Versus Total Intergenerational Transfers. In addition to redefining life cycle wealth, Modigliani redefines intergenerational transfers to include only bequests at death. Thus, his flow calculations of the shares of life cycle and transfer wealth exclude all intergenerational transfers in the form of explicit gifts, college tuition, and implicit gifts such as interest free loans, the transfers of businesses to children though partnership agreements, and so on. In effect, Modigliani treats any adult, regardless of age, who received non-bequest transfers from his or her parents, as a "dependent" and ascribes the consumption resulting from such transfers to the parent(s).

In contrast to this peculiar redefinition of intergenerational transfers and adulthood, Kotlikoff and Summers (1981) defines the age of adulthood as age 18 and treats all payments, either in cash or in kind (including tuition payments), received from parents by children above age 18 as an intergenerational transfer. Support of children prior to age 18 is considered consumption by the parent. Because the consumption of children is ascribed to their parents, the choice of age of adulthood affects the shape of the age-consumption profile of adult cohorts.

Modigliani particularly objects to our labelling college tuition expenditures by parents as an intergenerational transfer. While one may argue about the proper age of adulthood, given the fungibility of money there is no reasonable basis for labeling parental tuition support differently from parental gifts of durables, such as cars, or parental gifts of money; that is, whether the parent pays tuition or gives the child the money to pay tuition is economically equivalent. In addition, there is no reason, as

Modigliani suggests, to classify somehow educational expenditures as a human as opposed to nonhuman wealth transfer. The transfer of funds to pay for education constitutes a transfer of nonhuman capital. The fact that the expenditure leads to smarter or more skilled children as opposed, for example, to fatter children, is quite immaterial to the issue of tracing the origins of nonhuman wealth accumulation.

The choice of the age of adulthood is, however, arbitrary. While age 18 seems reasonable for the postwar period, it may be too old for older 1974 cohorts, some of whom were born in the last century. Many of these older generations entered the labor force at younger ages than is currently typical and they certainly had a much shorter lifespan. Indeed, until the 1950s labor force participation rates were calculated for the population over age 14. Had Summers and I used a younger age, such as 16, for the age of adulthood for the older cohorts that were alive in 1974, we would have reported considerably less life cycle wealth.

Simulation Studies of Life Cycle Wealth and Comparisons with Actual Wealth Holdings

Simulation analyses also call the pure life cycle model into question. Auerbach and Kotlikoff (1985, 1987) point out, using a detailed life cycle simulation model, that realistic specification of U.S. demographics, preferences, and fiscal institutions implies an extremely small, if not negative, wealth to income ratio. These results differ from those of Tobin (1967) because of their inclusion of income taxes and social security and their more realistic assumptions concerning the growth rate of consumption over the life cycle. To generate substantial life cycle savings, Tobin found it necessary to assume that consumption over the life cycle grows at a much faster rate than actually observed. Tobin's calculations which appear to come closest to replicating observed ratios of wealth to income assume that consumption grows at a rate of 5 percent per year over the life cycle. This rate is more than twice the rate actually observed (Kotlikoff and Summers, 1981).

Other simulation studies by Atkinson (1971) and Oulton (1976) point out the difficulty of explaining wealth inequality based on the zero transfer life cycle model. To quote Atkinson, "It is clear from the analysis that life cycle factors cannot explain the upper tail of the current distribution of wealth in Britain." Note that the upper tail of the British wealth distribution in Atkinson's study accounts for about three-quarters of total British wealth holdings. Both Atkinson and Oulton indicate that the substantial inequality in wealth relative to earnings can only be explained by intergenerational transfers. The Bevan and Stiglitz (1979) simulation study explicitly accounted for "altruistic" bequests, but just for lower income children, and showed that even this could not account for the tail. Finally, it appears that even adding the assumption of unintended bequests arising from imperfect annuity markets to the life cycle model cannot explain the tail.

Asset Decumulation of the Elderly

Decumulation of wealth after retirement is an essential aspect of the life cycle theory. Yet simple cross section tabulations of wealth holdings by age (Mirer, 1979;

Kurz, 1984a) or savings rates by age (Thurow, 1976; Danziger *et al.*, 1982–83) do not support the central prediction that the aged dissave. Mirer and Kurz report that wealth holding tends to increase with age.

Thurow reports positive saving rates for persons in all age groups, while Danziger *et al.* report that saving rates increase with age with "the elderly spending less than the nonelderly at the same level of income and (with) the very oldest of the elderly have the lowest average propensity to consume." The most recent cross section study of consumption rates by age is that of Albert Ando (1986a); his regression analysis leads him to conclude that "most families follow a reasonably well defined pattern of savings and net wealth accumulation before their retirement, but they tend to dissave little after retirement."

Cross section analysis of estate data by Atkinson (1971) and Atkinson and Harrison (1978) as well as panel analysis of the estates of individual cohorts by Menchik and David (1983) and David and Menchik (1985) also show that the average amount of wealth of cohorts rises in old age. Shorrocks' (1975) panel study of the estates of a single cohort in England indicates positive wealth accumulation until age 70. Hayashi's (1986) analysis of savings of Japanese cohorts in the 1970s leads him to conclude that "mean asset holdings do not decline as cohorts age," and Ando (1986b) commenting on Hayashi's research on Japan as well as his own states that "the apparent total lack of dissaving by older households in Japan is clearly inconsistent with the life-cycle theory." Menchik and David's longitudinal analyses of American estate data is most relevant for understanding U.S. savings behavior. They examine the estates of four birth cohorts living in Wisconsin in the period 1947 to 1978 and born between 1880 and 1925. In addition to studying within cohort behavior, their analyses control for the lifetime earnings of cohort members. Their results "fail to show individuals decumulating wealth in old age."

In contrast to these studies, the study of King and Dicks-Mireaux (1982), based on cross section Canadian data, suggests a positive rate of decumulation in old age. Burbridge and Robb (1985), on the other hand, reach a somewhat different conclusion from an examination of King and Dicks-Mireaux's data. They report old age asset decumulation only for blue collar workers. For white collar workers the age wealth profile is flat between age 55 and 65 and rises after age 65. Except for the study of King and Dicks-Mireaux, the only other analysis that seems to confirm the life cycle model's prediction of a declining age-wealth profile is the panel regression analysis of Diamond and Hausman (1984). However, as both Bernheim (1986) and Hurd (1986) point out, their subsample from the National Longitudianal Survey may not be representative. Indeed, Diamond and Hausman begin their analysis by excluding low wealth households on the grounds that their behavior is inconsistent with the assumptions of the life cycle model. In addition, their data set is not well suited to a study of the wealth of the elderly after they retire because even at the end of the 10-year panel the ages of the sample range from 55 to only 69.

A number of questions can be raised about many of these analyses of the age wealth profiles and saving rates of the aged. The cross section and several of the cohort analyses (but not those of Menchik and David) may be biased towards a slower rate of asset decumulation because the rich live longer than the poor. In addition, the composition of the sample by age of retirement changes with the age of the sample. But these findings concerning the failure to decumulate are particularly striking since, as Bernheim (1986) points out, the presence of future social security and pension streams increases the rate of decumulation of other assets implied by the life cycle model.

Probably the best analyses of asset changes among the aged are those of Hurd (1986) and Bernheim (1986) both of which use the panel data in the Retirement History Survey. While panel analysis permits holding constant the individual household, one cannot be sure whether changes in wealth over time reflect conscious household saving decisions or are merely the result of unexpected capital gains or losses.

Bernheim considers change in wealth over two periods, 1969–75 and 1975–79, for two samples of elderly, the retired and the nonretired. He also distinguishes between couples and single individuals. Bernheim's measure of wealth change is the log of the ratio of wealth at the end of the period to wealth at the beginning of the period.

For retired couples Bernheim reports a quite small average rate of asset decumulation (a small average ratio of the logs) for the first period and a small average rate of asset accumulation in the second period. Almost half of retired couples in the earlier period and almost three-fifths of couples in the latter period exhibited positive rates of asset accumulation. For nonretired couples the average rate of wealth change was positive in both periods with roughly three-fifths of the nonretired elderly engaging in positive accumulation. Nonretired single individuals also had positive accumulation on average in both periods, again with about three-fifths exhibiting positive saving. The only group whose average rate of wealth change was negative in both periods is retired single individuals. This group displayed a reasonably large rate of asset decumulation on average; but even among this group over two-fifths had positive saving.

Bernheim carefully considers these data in light of social security and private pensions and states that (1) "the inclusion of annuities reinforces earlier findings that resources decline only slightly, if at all, after retirement;" and (2) "the pure life cycle hypothesis fails to account for saving behavior after retirement."

Hurd reaches a different conclusion from Bernheim, stating "contrary to many results from cross-section data, the elderly do dissave." However, this conclusion appears to be based on tabulations which exclude housing wealth. When Hurd includes housing wealth, he reports that average total real wealth of all Retirement History Survey couples declined by only 2 percent between 1969 and 1979. For single individuals Hurd reports a decline of 22.4 percent over the 10 year period in average total wealth. The corresponding figures when housing wealth is excluded are 14.5 percent for couples and 36.4 percent for singles.

Hurd views the findings on non-housing wealth as most informative, arguing that increases in housing wealth in the 1970s reflect unexpected capital gains. He fails to point out, however, that the reduction in non-housing wealth over the period may reflect unexpected capital losses on stocks and bonds. In addition, the elderly may well have responded to capital gains in housing by accumulating less in other forms; that is, in the absence of capital gains in housing, non-housing wealth may have been larger.

Hurd's measure may also be more sensitive to measurement error than Bernheim's; the level of average wealth in either 1969 or 1979 may be biased because of outliers. Hurd has indicated (to the author) that the results based on median wealth are similar, but his study, in contrast to that of Bernheim, tells us little about the distribution of wealth changes across the sample.

In sum, the panel studies of Bernheim and Hurd suggest little if any total wealth decumulation of couples and some limited total wealth decumulation of single individuals. In addition, Bernheim shows that a sizeable fraction of singles as well as couples continues accumulating wealth in old age. Even those panel studies which suggest wealth decumulation, on average, indicate a significant fraction of elderly households continuing to accumulate wealth. This, by itself, is strong evidence against the standard, no intergenerational transfer, life cycle model.

Evidence From Annuity Markets

The strict life cycle model without allowance for bequest motives makes strong predictions about the demand for annuities. Since the date of death is uncertain and since the bequests provide no utility, life cycle models imply that there should be a very strong demand for annuity insurance. Indeed, Kotlikoff and Spivak (1981) demonstrate that, for the commonly used time separable CES utility function of consumption and assuming a risk aversion coefficient of 1.75, the gain to a 55 year old with average mortality probabilities from obtaining access to an actuarially fair annuities market is equivalent to an almost 70 percent increase in his lifetime resources in the absence of such markets.

In fact, the demand for annuities appears to be very weak (J. Friedman and Spivak, 1986). Friedman and Warshawsky (1985) report that the loads on annuity insurance are no higher than the loads on other frequently purchased types of insurance such as property and casualty insurance. Yet annuity purchases are a rarity. Friedman and Warshawsky argue that it is necessary to assume a fairly strong bequest motive to explain this behavior. Friedman and Warshawsky do not, however, take into account uncertainty with respect to inflation risk, which may explain the reluctance of many to purchase nominal annuities. In addition, totally annuitizing one's wealth might leave one illiquid and unable to pay major one time expenses such as medical care. Still, the virtual nonexistence of annuity insurance is quite surprising.

Kotlikoff and Spivak (1981) advance a possible alternative explanation, namely that families will self-insure to a large extent when annuity insurance is only available on very unfavorable terms; family insurance, however, does not appear capable of fully accounting for the widespread failure to annuitize. Bernheim, Shleifer, and Summers (1985) review a number of settings where annuities are available on a fair or even subsidized basis and report that even in these cases there is little demand for annuities. They conclude from this evidence that many consumers must have significant bequest motives. Their strongest evidence comes from TIAA-CREF's 1973 Survey of Beneficiaries. In this survey over 70 percent of beneficiaries chose plans other than those providing full annuity protection despite the fact that TIAA-CREF annuities appear to be quite close to actuarially fair.

Historical Evidence

The essential prediction of the life cycle theory is that people save to prepare for their retirement when they must dissave and consume. Without periods of retirement, or a least, significant decreased labor earnings at the end of life there is no life cycle motive for saving. Yet there were substantial positive net national saving rates (net national saving divided by net national product) over the period 1870 to 1930 when retirement was much less common than it is today.

There was some retirement in the latter part of the last century and the pre-Depression years of this century, but the retirement rate of the elderly was small, especially when compared with current rates. During the period 1870 to 1930 the labor force participation rate of males 65 and older exceeded 58 percent and for much of the period it exceeded 70 percent (Moen, 1987a, 1987b). In contrast, the comparable current rate for males 65 and older is 25 percent. Life expectancy for 20 year-old males during the period 1870 to 1930 appears to have been about 65; for much of the postwar period it has been above 70.

Despite this significant increase in the length of retirement, the saving rate in recent decades has been substantially smaller than that observed between 1870 and 1930. The average saving rate, defined as net national saving divided by net national product, recorded from 1870 to 1930 is close to 50 percent greater than the average saving rate in the postwar period.[2] In contrast to the life cycle model, at least one class of bequests models suggests that saving rates may fall with the life span and, ceteris paribus, length of retirement (Skinner, 1985). The intuitive reason is that a reduction in morality probabilities reduces the likelihood of a near term bequest. Hence, there is less of an incentive to save for bequests. Of course, much besides the length of retirement changed over the last century, so the negative correlation between the saving rate and the length of retirement may be spurious. It is, therefore, also useful to examine this correlation over shorter periods. During the period 1869 to 1891 the labor force participation rate of men 65 and older exceeded 73 percent and at times was as high as 78 percent. In contrast, the participation rate between 1902 and 1931 was below 66 percent, and for most of the period it was below 60 percent. Despite this increase in the length of retirement over the two periods, the average rate of net national saving is slightly higher in the later period. There is also a postwar negative correlation between the length of retirement and the saving rate. Kotlikoff and Smith (1983) report that since 1950 the expected duration of retirement and other nonworking periods for the average adult has almost doubled. This change coincided with a secular decline since 1950 of almost 40 percent in the net national saving rate. Of course, the introduction of unfunded social security (Feldstein, 1974) and other

[2] *Historical Statistics*, p. 231, and *1987 Economic Report of the President*.

government policies may have reduced saving since 1950, but the life cycle model would still predict a quite substantial offsetting impact arising from the change in retirement (Kotlikoff, 1979).

Understanding Intergenerational Transfers

Explanations for Intergenerational Transfers

Altruistic concern for one's children is the first reason one thinks of for intergenerational transfers. This concern may be expressed mathematically as the parent having direct utility for the utility of the child as in Barro (1974) and Becker (1974). An alternative, nonaltruistic model of bequests (Yaari, 1966; Blinder, 1973) is that parents care about the level of the bequest per se. But a model where utility depends on the size of the bequest seems implausible because it implies that the parent's transfer to the child is independent of the child's economic circumstances.

The altruistic model also seems somewhat implausible because it has such strong neutrality predictions. As is well known, Barro (1974) showed that intergenerational transfers by the government are completely neutralized because parents internalize the effects on all future generations of such transfers. More recently, Kotlikoff (1983) and Bernheim and Bagwell (1986) independently showed that marriage may altruistically link large numbers of parents and other extended family members (if not the entire planet), with the implication that intragenerational government transfers will also be entirely neutralized.

The neutrality properties of the altruistic models hold only if those making transfers are not considered. As an example of such constraints, suppose an altruistic parent would like to transfer from her children to herself but the parent cannot compel her children to make such transfers; in this situation, government redistribution from that parent to her child will have real effects.

An alternative view of transfers from parents to children, which has nothing to do with parental concern for children or a desire per se to leave bequests, is that parents and children form an incomplete annuities market (Kotlikoff and Spivak, 1981). According to this view parents and children enter into implicit risk sharing arrangements. The risk to be shared is the parent's longevity risk. The arrangement involves parents agreeing to transfer their assets to their children at death as a quid pro quo for support payments from the children if the parents live longer than expected. Enforcement of such arrangements is enhanced by having the child make payments to the parent each year, rather than wait until the parent runs out of resources.

Kotlikoff, Shoven, and Spivak (1986, 1987) demonstrate that significant intergenerational transfers could arise in the aggregate in a dynamic equilibrium model of imperfect family annuity insurance. Indeed, they point out that such a model could explain much of the 80–20 split of transfer and life cycle wealth reported by Kotlikoff and Summers (1981). This point is important; it indicates that although intergenerational transfers may be large in the aggregate, fundamental preferences may still correspond to those posited in the generationally selfish life cycle model.

Other researchers, including Sheshinski and Weiss (1981), Davies (1981), Eckstein, Eichenbaum, and Peled (1983), Abel (1983), and Hubbard (1987), have examined the potential for intergenerational transfers when no annuity insurance whatsoever, including implicitly family insurance, is available. These studies also demonstrate clearly that significant intergenerational transfers can arise in the life cycle model if annuity insurance is unavailable.[3]

Even if the life span were certain, individuals may save significant sums for the possibility of substantial end-of-life medical and nursing home expenses (Kotlikoff, 1986). If such medical problems do not arise, the elderly individual may be unable to consume this nest egg before he or she dies, and, consequently, a bequest will arise.

Other individuals may simply strike it rich in their professions and investments and find themselves unable to consume all their resources. It's clear that consumption satiation does occur; there are enough examples of extremely wealthy individuals, many who are quite young, who are unable to consume their wealth over their lifetime, especially when one properly measures consumption as excluding expenditures on durables, but including imputed rent on durables. Additional research is needed to determine how substantial windfalls influence total wealth accumulation. While there is some contradictory evidence concerning the elasticity of how bequests change with respect to changes in resources (Menchik and David, 1983; Kotlikoff, 1977), this elasticity surely exceeds unity for the super-wealthy.

Bernheim, Shleifer, and Summers (1985) advance a fourth explanation for intergenerational transfers: the implicit payment by parents to their children for providing in kind services, such as assisting them with chores, trips to the doctor, and so on. While some intergenerational transfers surely arise as payments for child-provided services, it seems quite unlikely that such transfers would be very large in the aggregate.

Each of these explanations of intergenerational transfers presumably plays some role in explaining aggregate transfers; in addition, there are probably many traditional life cycle households that are well annuitized through private pensions and Social Security and, consequently, leave no intergenerational transfers. This heterogeneous view of preferences and insurance arrangements is espoused by Kurz (1984b).

Empirical Analyses of Alternative Models of Intergenerational Transfers

Empirical analysis of intergenerational transfers has been greatly limited by the available data. To test models of altruism or family insurance arrangements one needs data not only on the transfers, consumption choices, and resources of parents, but also those of adult children; that is, one needs information on the extended family. At the present time there simply are not suitable data sets covering the extended family.

There are, however, data on *inter vivos* transfers and bequests. In his analysis of 379 Connecticut estates with bequests to children with siblings Menchik (1980) finds strong evidence that "wealth bequeathed to children is shared equally." This study confirmed earlier findings by Sussman, *et al.* (1970) and Brittain (1978). On the other hand, Tome's (1981) findings based on recall data from Cleveland contradict the equal bequest view; but in 1985 Menchik (1985) sampled actual probate records in Cleveland and found equal division in 93 percent of the cases. The finding of equal bequests strongly contradicts the prediction of the altruistic model which predicts that differences in bequests would compensate for differences in children's earnings capacities. Donald Cox (1987) reports another finding at odds with altruism, namely that transfers are positively related to the recipient's income level.

---

[3] I find these life cycle models somewhat unsatisfactory because they simply assume away annuity insurance despite the fact that life cycle agents will have very substantial demands for annuities. In addition, the bequests arising in these models are arbitrarily allocated to children despite the fact that the parents have no bequest motive.

Hurd also finds evidence against the utility of bequest model, and indirectly against the altruism model. Hurd points out that parents with bequest motives should consume their end-of-life resources at a slower rate than those with bequest motives. He then compares the rate of asset decumulation of the elderly who have children with those who do not. He finds no significant difference in the rate of asset decumulation among the two groups. A contrasting finding by Friedman and Spivak (1986) is that people with more children are less likely to purchase annuities.

Bernheim, Shleifer, and Summers (1985) report that in settings with two or more children, children of richer parents spend more time with their parents than children of poorer parents. In contrast, they find no significant impact of parental wealth on the visitation of only children. The authors view these findings as strong support for their model in which parents with two or more children credibly threaten to disinherit those children who are insufficiently attentive. These findings, while suggestive, must be viewed as preliminary; further analysis, taking account of childrens' own wealth position, is needed.

In addition to these micro analyses of intergenerational transfers, Boskin and Kotlikoff (1985) use time series data to test directly one of the key neutrality propositions of the altruism model, namely that aggregated consumption is independent of the age distribution of resources. Their model assumes extended families have identical Barro-type preferences and identical demographic structures. Under these assumptions aggregate household consumption depends simply on the present value of aggregate human wealth, plus national net worth, less the present value of future government consumption. It is independent of the age distribution of resources. Boskin and Kotlikoff show that given the appropriate resource variables, the age composition of income still significantly influences aggregate consumption. Thus they reject their admittedly stylized model of intergenerational altruism. In contrast, Abel and Kotlikoff (1987) using micro data are unable to reject the Barro model's implication that, controlling for demographics, consumption of different age cohorts moves together.

Conclusion

Like most good puzzles, the saving puzzle has a lot of pieces, many of which do not seem to fit together. It seems well-established that intergenerational transfers are sizeable and that the process leading to these transfers is responsible for a sizeable amount of wealth accumulation. The precise explanation for these transfers remains unclear.

Intergenerational altruism might appear the most likely candidate, but at least some stylized facts, such as the equal allocation of bequests among children, are strongly at odds with the altruism model. Other explanations involving imperfect insurance arrangements or payments for child services do not appear capable of explaining the substantial amounts of transfers actually observed. Sorting out the relative contributions of different models to intergenerational transfers and the precise role of intergenerational transfers in the process of wealth accumulation remains an intriguing and exciting enterprise.

■ *I thank Michael Hurd, Franco Modigliani, James Poterba, Joseph Stiglitz, Lawrence Summers and Timothy Taylor for helpful comments.*

References

Abel, Andrew, "Bequests and Social Security with Uncertain Lifetimes," *American Economic Review*, September 1985, *75:4*, 777–91.

Abel, Andrew, and Laurence Kotlikoff, "Does the Consumption of Different Age Cohorts Move Together? A New Non Parametric Test of Intergenerational Altruism," mimeo, 1987.

Adams, Bert, *Kinship in an Urban Setting*. Chicago: Markham Publishing Company, 1968.

Ando, Albert, "Selected Comments on Current Empirical Evidence For or Against the Life Cycle Theory." Paper presented at The International Seminar on the Life Cycle Theory, Paris, 1986a.

Ando, Albert, Comment on Fumio Hayashi's paper in *NBER Macroeconomics Annual 1986*. Cambridge: The MIT Press, 1986b.

Ando, Albert, and Franco Modigliani, "The Life Cycle Hypothesis of Saving: Aggregate Implications and Tests," *American Economic Review*, March 1963, *53:1*, 55–84.

Atkinson, A. B., "The Distribution of Wealth and the Individual Life Cycle," *Oxford Economic Papers*, 1971, *23:2*, 239–254.

Atkinson, A. B., and A. J. Harrison, *The Distribution of Wealth in Britain*. Cambridge: Cambridge University Press, 1978.

Auerbach, Alan J., and Laurence J. Kotlikoff, *Dynamic Fiscal Policy*. Cambridge: Cambridge University Press, 1987.

Auerbach, Alan J., and Laurence J. Kotlikoff, "Simulating Alternative Social Security Responses to the Demographic Transition," *National Tax Journal*, June 1985.

Barro, Robert J., "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *82*, 1095–1130.

Becker, Gary S., "A Theory of Social Interactions," *Journal of Political Economy*, November/December 1974, *82:6*, 1063–94.

Becker, Gary S., and Nigel Tomes, "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *Journal of Political Economy*, December 1979, *87:6*, 1153–89.

Becker, Gary S., and Nigel Tomes, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, August 1976, *84*, S143–S162.

Behrman, J., R. Pollak, and P. Taubman, "Parental Preferences and Provision for Progeny," *Journal of Political Economy*, February 1982, *90:1*, 52–73.

Bernheim, B. Douglas, "Dissaving After Retirement: Testing the Pure Life Cycle Hypothesis." In *Issues in Pension Economics*. NBER volume. Chicago: University of Chicago Press, 1986.

Bernheim, B. Douglas, and Kyle Bagwell, "Is Everything Neutral?" NBER Working Paper, 1986.

Bernheim, B. Douglas, Andrei Shleifer, and Lawrence H. Summers, "The Strategic Bequest

Motive," *Journal of Political Economy*, December 1985, *93:6*, 1045–1075.

Bevan, D. and Joseph Stiglitz, "Intergenerational Transfers and Inequality," *The Greek Economic Review*, August 1979, *1:1*, 8 + .

Blinder, Alan, "A Model of Inherited Wealth," *The Quarterly Journal of Economics*, 1973.

Boskin, Michael J., and Laurence J. Kotlikoff, "Public Debt and U.S. Saving: A New Test of the Neutrality Hypothesis." Carnegie-Rochester Conference Volume Series, February 1985.

Brittain, John A., *Inheritance and the Inequality of Material Wealth*. Washington, D.C.: The Brookings Institution, 1978.

Burbridge, J. B., and A. L. Robb, "Evidence on Wealth Age Profiles in Canadian Cross Section Data," *Canadian Journal of Economics*, November 1985, *58:4*, 854–875.

Cox, Donald, "Motives for Private Income Transfers," mimeo, 1987.

Danziger, S. et al., "The Life Cycle Hypothesis and the Consumption Behavior of the Elderly," *Journal of Post Keynesian Economics*, Winter 1982–83, *2*, 208–227.

Darby, Michael R., *Effects of Social Security on Income and the Capital Stock*. Washington, D.C.: American Enterprise Institute, 1979.

David, Martin, and Paul L. Menchik, "The Effect of Social Security on Lifetime Wealth Accumulation and Bequests," *Economica*, November 1985. *52:208*, 421–434.

Davies, James, "Uncertain Lifetime, Consumption, and Dissaving in Retirement," *Journal of Political Economy*, June 1981, *89:3*, 561–577.

Diamond, Peter A., and Jerry Hausman, "Individual Retirement and Savings Behavior," *Journal of Public Economics*, February/March 1984, *23:1/2*, 81–114.

Eckstein, Zvi, Marty Eichenbaum, and Dan Peled, "Uncertain Lifetimes and the Welfare Enhancing Properties of Annuity Markets and Social Security." mimeo, 1983.

Feldstein, Martin S., "Social Security, Induced Retirement, and Saving," *The Journal of Political Economy*, September/October 1974, *82:5*, 905–925.

Friedman Benjamin, and Mark Wahrshawsky, "The Cost of Annuities: Implications for Saving Behavior and Bequests." NBER Working Paper No. 1682. Cambridge: National Bureau of Economic Research, August 1985.

Friedman Joseph, and Avia Spivak, "Is There Adverse Selection in the Demand for Annuities?" Mimeo, 1986.

Hayashi, Fumio, "Why is Japan's Saving Rate So Apparently High?" *NBER Macroeconomics Annual* 1986. Cambridge: The MIT Press, 1986.

Hubbard, Glenn R., "Uncertain Lifetimes, Pensions, and Individual Saving" In Bodie, Z., J. B. Shoven, and D. A. Wise, eds. *Issues in Pension Economics*, Chicago: University of Chicago Press, 1987.

Hubbard, Glenn R., "Precautionary Saving Revisited: Social Security, Individual Welfare, and the Capital Stock," 1984b.

Hurd, Michael, "Savings and Bequests." NBER Working Paper No. 1826. Cambridge: National Bureau of Economic Research, January 1986.

Ishikawa, Tsuneo, "Family Structures and Family Values in the Theory of Income Distribution," *Journal of Political Economy*, October 1975, *83:5*, 987–1008.

King, Mervyn, and Louis Dicks-Mireaux, "Asset Holding and the Life Cycle," *The Economical Journal*, 1982, *92*, 247–67.

Kotlikoff, Laurence J., "Estimating the Elasticity of Bequests with Respect to Lifetime Resources from a Sample of Potential Decedents." mimeo, 1977.

Kotlikoff, Laurence J., "Social Security and Equilibrium Capital Intensity," *The Quarterly Journal of Economics*, May 1979, *93*, 233–53.

Kotlikoff, Laurence J., "Some Economic Implications of Life Span Extension." In *Biology and Behavior of the Aged*, National Institute of Aging, Academic Press, 1982.

Kotlikoff, Laurence J., "Health Expenditures and Precautionary Savings," NBER Working Paper No. 2008, August 1986.

Kotlikoff, Laurence J., "Theoretical and Empirical Analysis of Altruistic Linkages Within the Extended Family." mimeo, June 1983.

Kotlikoff, Laurence J., and Daniel E. Smith, *Pensions in the American Economy*. Chicago: University of Chicago Press, 1983.

Kotlikoff, Laurence J., and Avia Spivak, "The Family as an Incomplete Annuities Market," *Journal of Political Economy*, April 1981, *89:2*, 372–91.

Kotlikoff, Laurence J., and Lawrence H. Summers, "The Contribution of Intergenerational Transfers to Total Wealth: A Reply." NBER Working Paper No. 1827, February 1986.

Kotlikoff, Laurence J., and Lawrence H. Summers, "The Role of Intergenerational Transfers in Aggregate Capital Accumulation." *Journal of Political Economy*, August 1981, *89:4*, 706–732.

Kotlikoff, Laurence J., John Shoven, and Avia Spivak, "Annuity Market's Savings, and the Capital Stock." In Bodie, Z., J. B. Shoven and D. A. Wise, eds., *Issues in Pension Economics*, Chicago: University of Chicago Press, 1987, 211–234.

Kotlikoff, Laurence J., John Shoven and Avia Spivak "The Impact of Annuity Insurance on

Savings and Inequality," *Journal of Labor Economics*, July 1986, *4:3*, part 2.

Kurz, Mordecai, "Capital Accumulation and the Characteristics of Private Intergenerational Transfers," *Economica*, February 1984a, *51:201*, 1–22.

Kurz, Mordecai, "The Life Cycle Hypothesis as a Tool of Theory and Practice." mimeo, 1984b.

Loury, Glen, "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, July 1981, *49:4*, 843–867.

Menchik, Paul L., "Primogeniture, Equal Sharing, and the U.S. Distribution of Wealth," *Quarterly Journal of Economics*, March 1980, *94:2*, 299–316.

Menchik, Paul L., "Unequal Estate Division: Is it Altruism, Reverse Bequests, or Simply Noise?" Paper presented to the 1984 Seminar on Modeling the Accumulation and Distribution of Wealth, Paris, France, 1985.

Menchik, Paul, and M. David, "Income Distribution, Life Time Saving and Bequests," *American Economic Review*, September 1983, *83:4*, 672–690.

Mirer, Thad W., "The Wealth Age Relation Among the Aged," *American Economic Review*, June 1979, *69:3*, 435–443.

Modigliani, Franco, "The Contribution of Intergenerational Transfer to Total Wealth." Paper presented to the Paris Conference on Modeling the Accumulation and Distribution of Wealth, September 1984.

Modigliani, Franco, "The Life Cycle Hypothesis of Saving, Twenty Years Later." In Parkin, M., ed., *Contemporary Issues in Economics*. Manchester: Manchester University Press, 1975, pp. 2–36.

Modigliani, Franco, and Richard Brumberg, "Utility Analysis and Aggregate Consumption Functions: An Attempt at Intergation." In Abel, A., ed., *The Collected Papers of Franco Modigliani*. Cambridge. The MIT Press, 1980. Vol. *2*, 128–197.

Modigliani, Franco, and Richard Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data." In Kurihara, K. K., ed., *Post Keynesian Economics*. New Brunswick, N.J.: Rutgers University Press, 1954, pp. 388–436.

The President's Council of Economic Advisors, "The 1987 Economic Report of the President." Washington, D.C.: U.S. Government Printing Office, 1987.

Moen, John, "The Labor of Older Men: A Comment," *The Journal of Economic History*, September 1987a.

Moen, John, "Essays on the Labor Force and Labor Force Participation Rates: The United States from 1860 through 1950." Unpublished Ph.D. dissertation, University of Chicago, 1987b.

Oulton, Nicholas, "Inheritance and the Distribution of Wealth," *Oxford Economic Papers*, March 1976, *28*, 86–101.

Ransom, Roger L., and Richard Sutch, "The Labor of Older Americans: Retirement of Men On and Off the Job, 1870 to 1937," *The Journal of Economic History*, March 1986.

Sheshinski, Eytan, and Yoram Weiss, "Uncertainty and Optimal Social Security Systems," *Quarterly Journal of Economics*, May 1981, *94:2*, 189–206.

Shorrocks, Anthony F., "The Age Wealth Relationship: A Cross Section and Cohort Analysis," *Review of Economics and Statistics*, May 1975, *57:2*, 155–63.

Skinner, Jonathan, "The Effect of Increased Longevity on Capital Accumulation," *American Economic Review*, December 1985, *75:5*, 1143–1150.

Solow, Robert, "Reflections on Saving Behavior." In *Saving and Government Policy*. Conference Series No. 24. Boston: Federal Reserve Bank of Boston, 1982.

Summers, Lawrence H., "Capital Taxation and Capital Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, *71:4*, 533–544.

Sussman, M. B., J. N. Cates, and O. T. Smith, *Inheritance and the Family*. New York, Sage, 1970.

Thurow, Lester, *Generating Inequality*. New York: Basic Books, *1976*.

Tobin, James, "Life Cycle Saving and Balanced Growth." In Fellner, William *et al.*, eds., *Ten Economic Studies in the Tradition of Irving Fisher*. New York: Wiley, 1967, pp. 231–56.

Tomes, Nigel, "The Family, Inheritance, and the Intergenerational Transmission of Inequality," *Journal of Political Economy*, October 1981, *89:5*, 928–58.

U.S. Department of Commerce, Bureau of Census, *Historical Statistics of the United States*. Washington D.C.: U.S. Government Printing Office, 1975.

White, Betsy Buttrill, "Empirical Tests of the Life Cycle Hypothesis," *American Economic Review*, September 1978, *68(4)*, 547–60.

Yaari, Menachim E., "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Review of Economic Studies*, April 1965, *32:2*, 137–150.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 59–62

Symposium on Public and Private Unionization

Edward P. Lazear

The past few years have witnessed a divergence between rates of unionization in the private economy and the public sector. Union membership as a proportion of total employment has declined significantly in the private sector, while unionization of workers in the public sector has increased rather significantly. Although valid for the United States, the pattern does not hold internationally. A number of authors have pointed out that Canada, for example, which has a similar industrial structure to the United States (indeed, many of the same firms), has experienced an increase in unionization in the private sector. Even in the United Kingdom, where rates have declined recently, unionization grew during the 1970s. In the two papers that follow, the authors address the changes in unionization and attempt to explain the changes.

Melvin Reder argues that all time periods cannot be explained by the same mechanism. His explanation of private sector unionism emphasizes general economic conditions. He argues that when economic conditions are favorable and unemployment rates are low, unionization increases. Stagnant economies and declining ones are characterized by reductions in unionization. He admits to being unable to explain the experience of the 1930s in this way, and resorts to other explanations. Richard Freeman, on the other hand, explains the recent decline in unionization rates by arguing that employers have taken a more aggressive stand against unions. One of the best pieces of evidence for this hypothesis, Freeman claims, is the enormous increase in antiunion activity as measured by NLRB data on unfair labor practices. He discusses empirical evidence that suggests that the extent of management opposition substantially determines the outcomes of organizing campaigns. It is more difficult for the

■ *Edward P. Lazear is Isadore Brown and Gladys J. Brown Professor of Urban and Labor Economics, University of Chicago, Chicago, Illinois, and Senior Fellow at The Hoover Institution, Stanford University, Stanford, California.*

Freeman hypothesis to explain why management's anti union activity has increased over time and why it was not a force that was present in earlier periods.

Both authors agree that the role of laws and government regulations is large in explaining the rise in unionization in the public sector. Thus, the theories are bifurcated, and there is no common explanation for the decline of unionization in the private sector and the increase in the public sector.

As interesting as it is to explain the decline in unionization, it is also important to ask whether the changes are of any consequence. Suppose that the trend continues and private-sector unionization declines. Suppose further that unionization of the public sector increases. What will that mean to the nature of industrial relations in this country and, more generally, to the average level of productivity? Obviously that question is too big to answer in a few paragraphs, but since neither author addresses these issues, I outline some of what is at stake.

First, it is reasonably well established that unions do create a positive wage differential between the union and nonunion sector. Even if wages were not raised on average as a result of organizing activity, the higher wages in the union sector cause a substitution effect: too much is produced by the nonunion firms and too little is produced by union firms. Additionally, union firms tend to have too high a capital/labor ratio, to mitigate the higher labor costs imposed by having to pay union wage rates. If unionization is not neutral across the economy, then too few union products are sold relative to nonunion products. A decline in unionization rates can reverse this phenomenon as long as two things are true. First, the decline must not be so lopsided so that interproduct distortions are increased rather than decreased as a result of a decline. Second, the decline in unionization must not occur because wage differentials have become even more pronounced. In particular, if the decline in private-sector unionization is the result of increased wage pressure by unions, which effectively drives union firms under relative to nonunion firms, then distortions may get worse rather than better as one observes union rates declining. This point bears some clarification.

If Reder is correct and changes in private sector unionization result from differences in overall economic conditions, then reductions in unionization rates will be associated with lower wage differentials between union and nonunion workers. Economic distortions will decline as a result of the fall in unionization rates. But if Freeman is correct, it is necessary to inquire why anti union pressure has increased. One possibility is that unions have become more aggressive in seeking wage gains. If the unions are successful in raising wage differentials, then distortions may get worse and the decline in unionization would be merely one reflection of that phenomenon.

There is one wrinkle, a point that was made many years ago and that has become even clearer in the context of the contract literature. Higher wages do not necessarily lead to a distortion in factor utilization rates. As long as unions can bargain over employment as well as wages, it is possible that the outcome of the bargaining process results in having the efficient amount of labor employed. Since points on the efficiency frontier dominate inefficient points, if a sufficiently rich set of transfer mechanisms exists, then union workers and firms can be made better off by adopting efficient factor choices. It is much less likely that efficiency is obtained in the product markets, however. For this to ensue, a central planner would have to allocate goods independent of prices.

The point suggests that changes in unionization rates have different effects in the private economy than they do in the public sector. First of all, the public sector faces little or no competitive pressure in the product market. Even those goods that do have their private-sector counterparts need not be priced by the public sector so as to cover costs of production. For example, a city may offer a recreational facility at a price that is competitive with the one charged by a private club. Even if the city is at a cost disadvantage, the city makes up the deficit by levying taxes on its residents. Under certain circumstances, no distortion is created and the city merely acts as an agent that transfers money from its citizens as a whole to city employees. Public schools that must compete with private schools are perhaps a more important example. School districts rely on general sources of funds, rather than student tuition, to support education that is provided at zero price. But teachers must be paid as much, often more, than those who teach at private schools.

Even in factor markets the public sector can act in a noncompetitive way. Although governments must buy their labor in competitive labor markets, the lack of a profit constraint allows the government to adopt any factor mix that it chooses. Again, the government could, by choice, adopt the efficient factor mix and merely decide to transfer resources from the public at large to government workers. Thus the growth of unionization in the public sector could, under ideal circumstances, imply no distortion and only transfers from one group of citizens to another.

Unions have effects on variables other than wages. At one extreme are notorious cases of featherbedding, where unions are able to require that inefficient production techniques are used. Unions have also been charged with hindering the efficient assignment of workers by using rigid seniority structures and inflexible promotion rules. Additionally, unions generally oppose giving employers discretion to pay according to performance because, unions argue, the power is exercised in an arbitrary fashion. Whether the decline in unionization will improve heavily unionized industry's international competitiveness as a result of the loosening of these restrictions remains to be seen. As Freeman points out, declining unionization in the private sector cannot be explained by looking at changing industrial composition, which might occur as unions priced certain industries out of the competition. However, it may be still be true that, within a given industry, nonunionized firms expand and unionized ones contract in response to international competition.

On the other side of the coin are the arguments by some that unions actually improve the productivity of firms. Unions set up grievance procedures and other internal organization devices that may facilitate industrial relations. As such, they can be thought of as an extension of the firm's personnel department. To the extent that the loss of unionization is not compensated for by other institutions that perform the same functions, productivity could actually decline. Again, however, if one believes that efficient arrangements can be made in the labor market, it is difficult to accept that employers and workers would not voluntarily set up the alternatives, since both sides could be made better off. Union advocates argue that firms oppose unions because setting up the institution that makes the firm productive also implies a rent transfer from stockholders to union members.

Understanding the source of the decline of private sector unions and concomitant rise of public sectors unions is essential. Without that knowledge, the welfare effects of the changes cannot be ascertained. The two essays that follow take us a long way down the road toward that understanding.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 63–88

Contraction and Expansion: The Divergence of Private Sector and Public Sector Unionism in the United States

Richard B. Freeman

The institutional structure of the American labor market changed remarkably from the 1950s and 1960s to the 1980s. In the '50s and '60s trade unions seemed permanently established in the private sector of the economy: a third of nonagricultural wage and salary workers and over half of blue-collar workers were union members; hundreds of thousands of workers voted annually in National Labor Relations Board (NLRB) representation elections to join unions; most large firms sought stable collective bargaining relations with their unions. By contrast, in the public sector only 10–12 percent of workers were union members; fewer were covered by collective bargaining contracts; and most experts regarded public employees as intrinsically nonorganizable. According to AFL-CIO president George Meany, it was "impossible to bargain collectively with the government" (Kramer, 1962, p. 14).

The massive contraction of unionism in the private sector and expansion in the public sector in the 1970s and 1980s (see Exhibit 1) has produced an utterly different situation today. In the private sector, the proportion of wage and salary workers in unions plummeted to 14 percent in 1986—a level comparable to that in the Great Depression; only a minuscule number of workers joined unions through NLRB elections; and national companies openly proclaimed their intent to establish a "union-free environment." By contrast, in the public sector over a third of the work force was unionized; some 40 percent were covered by collective contracts;[1] and

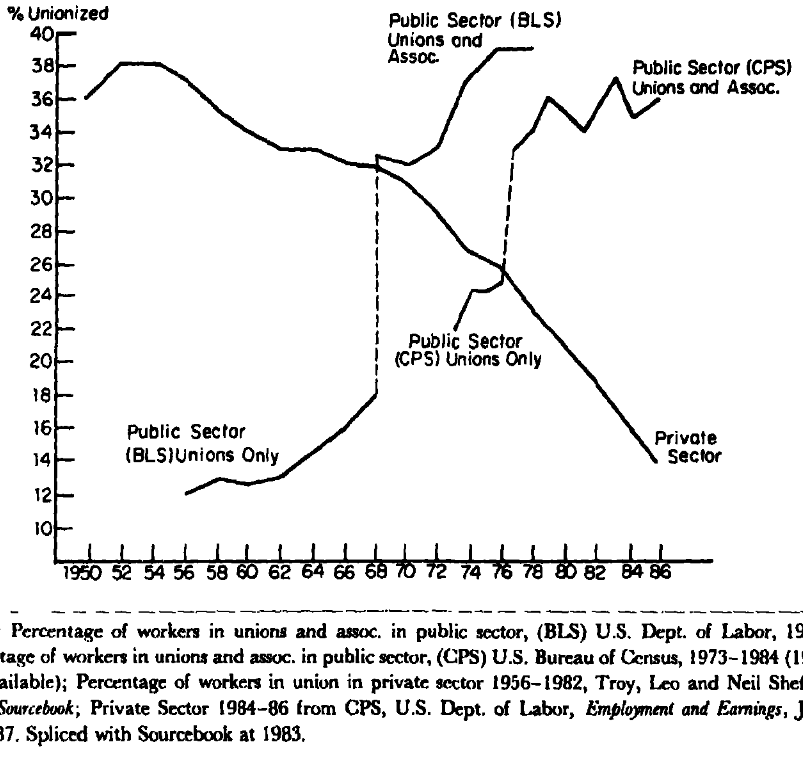
---

[1] Estimates of organization in the public sector differ somewhat among sources, though all data show greater organization than in the private sector. See Freeman, Ichniowski, and Zax (forthcoming) for a detailed analysis of the various statistics.

■ *Richard B. Freeman is Director of Labor Studies, the National Bureau of Economic Research, Cambridge, Massachusetts, and Professor of Economics, Harvard University, Cambridge, Massachusetts.*

Exhibit 1

Changing percent of non-agricultural workers who are union, by sector.



Sources: Percentage of workers in unions and assoc. in public sector, (BLS) U.S. Dept. of Labor, 1979; Percentage of workers in unions and assoc. in public sector, (CPS) U.S. Bureau of Census, 1973–1984 (1982 not available); Percentage of workers in union in private sector 1956–1982, Troy, Leo and Neil Sheflin, *Union Sourcebook*; Private Sector 1984–86 from CPS, U.S. Dept. of Labor, *Employment and Earnings*, Jan. 1985, 87. Spliced with Sourcebook at 1983.

public sector unions such as the American Federation of State, County and Municipal Employees, National Education Association, Service Employees International (which became largely public sector in the period), and the American Federation of Teachers were among the largest in the country. With one in three union members working in the public sector, women with master's degrees (largely schoolteachers) more highly organized than male high school graduates, and police and firefighters the exemplars of craft unionism, the union movement differed drastically from that headed by Meany in the 1960s.

Because the private sector employs nearly 85 percent of nonagricultural wage and salary workers in the United States,[2] the contraction of unionization in the private sector dominates the trends for the economy as a whole, with the result that the union proportion of nonagricultural employees fell from 36 percent in 1956 to 18 percent in 1986. This decline in union density was larger than that of the 1920s, and thus arguably represents the most significant change in labor market institutions since the Depression—the effective de-unionization of most of the U.S. labor force.

---

[2] U.S. employment data come from two basic sources: Current Population Survey data on individuals and establishment data from the Bureau of Labor Statistics. The 85 percent figures comes from the establishment survey.

What explains the decline in union representation of private wage and salary workers? Why have unions expanded in the public sector while contracting in the private sector? Is the economy-wide fall in density a phenomenon common to developed capitalist economies, or is it unique to the United States? To what extent should economists alter their views about what unions do to the economy in light of the fact that they increasingly do it in the public sector?

To answer these questions I examine a wide variety of evidence on the union status of public and private workers. I contrast trends in unionization in the United States with trends in other developed countries, particularly Canada, and use these contrasts and the divergence between unions in the public and private sectors of the United States to evaluate proposed explanations.

Because standard economic theory is not normally used to analyze massive changes in the institutional structure of an economy, many may think that an explanation of changing union density requires extensive reliance on social factors that go beyond economics. I argue that this is not the case and base my analysis on economic forces and behavior. The complexity of institutional change does, however, dictate a catholic analytic approach, and I use several lines of argument and evidence rather than a single econometric model or hypothesis test to reach conclusions. In addition, I recognize that a significant residual remains that can and perhaps should be attributed to noneconomic factors such as the abilities of union leaders, public opinion, and the like.

The Dimensions of Change

> "*It is a capital mistake to theorize in advance of the facts.*"
> Arthur Conan Doyle, recording the words of S. Holmes

Three firm observations can be made about the change in union density in the United States: first, that the decline in density in the private sector has been virtually ubiquitous, encompassing workers in all industries, regions, and occupations during the 1980s and earlier; second, that by contrast union expansion in the public sector has been highly uneven, occurring rapidly in some states and occupations but not in others; and third, that the overall drop in union density—the de-unionization of the economy—is distinctly American, contrasting sharply with developments in most Western countries.

Exhibit 2 documents the claim that union density has fallen among virtually all workers during the 1980s, when the rate of decline was exceptionally severe. The industry figures show the union proportion of workers falling in the areas of traditional union strength such as transportation and public utilities (including trucking, which the Teamsters once dominated); construction (long the preserve of craft

Exhibit 2

Percentage of private wage and salary workers who are members of unions, by industry and occupation 1980–1986

	% Organized	
	1980	*1986*
Industry		
Mining	32	14
Construction	31	22
Manufacturing	32	24
Transportation, communication, and public utilities	48	35
Trade	10	7
Service	9	6
Occupation		
Professional tech & kindred	23	19
Managers and administrators	8	7
Clerical and kindred	16	14
Sales	4	6
Craft and kindred	39	29
Operatives, except transport	40	33
Transport & equipment	45	31
Non-farm laborers	33	25
Service workers, except protective service[a]	13	10

[a] Protective service excluded because they are largely in the public sector.

Occupations based on 1980 titles with 1985 estimates for:

Professional, tech & kindred = professional Specialty + Tech and related support, weighted by employment;

Managers & administration = executive, administrator and managers;

Clerical & kindred = precision production, craft & repair;

Operatives, except transport = Machine operators, assemblers & inspectors;

Transport Equipment = transportation & material moving occupations;

Non-farm laborers = handlers, equipment cleaners, helpers & laborers

Source: U.S. Department of Labor, 1980 and 1987.

unions); manufacturing (where industrial unions have held sway since the organizing drives of the 1930s and 1940s); and services and trade (where proportions organized were so low that even a modicum of union success would raise the union shares.) The occupation data (which understate declines in private sector density because they include public sector labor) reveal a similar trend, with sizeable drops in density among operatives, craft workers, and laborers. As for the longer run, the percentage of production workers in metropolitan areas with collective bargaining contracts fell from 73 percent in 1960–61 to 51 percent in 1984, with massive declines in every region of the country ranging from −16 points in the South (48 percent to 32 percent) to −32 points in the West (80 percent to 48 percent, Goldfield, 1987). These declines show that much more is involved in the contraction of union density in the private sector than changes in the regional mix or production worker share of employment.

Overall, the data indicate that the explanation of the decline in private sector union density should focus on factors that affect all private employees as opposed to factors that affect employees in some segments of the labor market rather than others.

The Rise in Public Sector Union Density

Measuring changes in unionism in the U.S. public sector is difficult. One difficulty is that each American state (and the federal government) regulates public sector labor in its own jurisdiction, producing labor organizations with different legal rights. Whereas in the private sector union membership, collective bargaining representation, and the right to strike are coterminous under the same legal code, in the public sector a worker can be a union member but not be covered by a collective contract in some states (because management can refuse to bargain or to sign an agreement with a union) while being covered in others; be covered by a contract that does not include wages and salaries, as in federal employment; be allowed to strike in some jurisdictions but obligated to resolve disputes through arbitration elsewhere, and so on. Historically, moreover, public sector worker organizations have ranged from associations opposed to collective bargaining to acting as aggressive unions, with the same organization serving as a union in some localities but not in others (the American Association of University Professors is a case in point). Indicative of the measurement problem, the figures on public sector density in Exhibit 1 show discontinuities when the relevant surveys first included "associations" with unions in their definition of collective organization. While it might seem misleading to compare union membership in early years with union plus association membership in later years, this comparison does in fact accurately measure the magnitude (though not the timing) of change. This is because in the 1950s and early 1960s most associations—for instance, the NEA or the various police and firefighters organization—did not engage in collective bargaining, often rejecting it as "unprofessional," whereas in the 1980s they embraced bargaining and the other attributes of traditional trade unions.

A final complexity is that the legal environment under which public sector unions operate has changed over time. From the 1960s through the early 1970s, when public sector workers joined unions or saw their associations turn into unions, most public sector labor laws did not require employers to bargain/resolve impasses and outlawed the strike weapon that unions use to force employers to come to agreement. As a result many workers who joined unions did not work under collectively bargained contracts. By contrast, from the mid-1970s to the 1980s more and more union workers (and nonunion workers in organized workplaces) were able to gain contracts from their employers (35 percent of state and local government workers were covered in 1982 compared to 26 percent in 1975).[3] At the same time that contract coverage was increasing, however, the proportion of public sector workers in unions stabilized or

---

[3] These data are from the Survey of Governments, which differ from those in the Current Population Survey, as described in Freeman, Ichniowski and Zax (forthcoming).

Exhibit 3

Geographic variation in percentage of employees represented by bargaining units, by occupation and state law, 1982

State laws from most favorable to least favorable to collective bargaining	*No. of states with given labor law (% employees represented by bargaining unit), by occupation:*			
	Police	*Fire*	*Sanitation*	*Street & Highway*
Arbitration or strike Permitted	15 (94%)	18 (94%)	13 (39%)	13 (49%)
Duty to bargain	12 (58%)	11 (62%)	11 (14%)	11 (71%)
Bargaining permitted	11 (45%)	13 (54%)	12 (23%)	12 (19%)
No provisions	8 (29%)	5 (30%)	9 (27%)	9 (44%)
Bargaining prohibited	4 (16%)	3 (18%)	5 (8%)	5 (1%)

Sources: U.S. Bureau of the Census, Survey of Governments, 1982 NBER State Public Sector Collective Bargaining Law Data Set.

fell. In this case, statistics on membership and coverage tell different stories about the trend in union strength. As collective bargaining resulting in a contract has been the essence of American union activity, I stress increased contract convergence rather than membership trends in assessing the 1980s.

Given differences in state regulations of public sector collective bargaining, it is perhaps not surprising to find that the extent and, by extension, expansion of bargaining has varied across states and among occupations (see Exhibit 3). In states with laws favorable to collective bargaining (arbitration or strike permitted laws, and duty to bargain), proportionately more workers are represented by bargaining units than in states that permit but do not require employers to bargain, while representation by unions is low in states that have no collective bargaining provisions and lowest in those that prohibit bargaining. Though these data relate to cross-section differences at a point in time rather than changes over time, we can infer patterns of growth from them. We can do this because contract coverage was sufficiently slight in the earlier period that states/occupations with high density in the 1980s almost always had rapid growth whereas those in which density was low could not have had such growth. Interpreted in this way, the exhibit shows the uneven expansion of union representation, with proportionately more workers gaining representation in states (and occupations) having laws favorable to collective bargaining than in other states.

The evidence on geographic divergence implies that an explanation of union growth in the public sector must have a geographic dimension related to state laws as opposed to the "across-the-boards" explanation that seems appropriate for the private sector.

Cross-Country Contrasts

Because the meaning and measurement of union membership differs across countries, the figures in Exhibit 4 should be regarded as crude indicators of degrees of

Exhibit 4

Levels and changes in union density as a percent of nonagricultural wage and salary across countries, 1970–85

	1970	*1979*	*1984/5*	*1970–79*	*1979–85*
Countries with sharp rises in density					
Denmark	66	86	98	+20	+12
Finland	56	84	85	+28	+1
Sweden	79	89	95	+10	+12
Belgium	66	77		+11	
Countries with moderate rises in density					
Italy (?)	39	51	45	+12	−6
Germany	37	42	42	+5	0
France (?)	22	28	28	+6	0
Switzerland	31	36	35	+5	−1
Canada	32	36	37	+4	+1
Australia	52	58	57	+6	−1
New Zealand	43	46	?	+3	
Ireland	44	49	51	+5	+2
Countries with stable / decline density					
Norway	59	60	61	+1	+1
United Kingdom	51	58	52	+7	−6
Austria[a]	64	59	61	−5	+2
Japan[a]	35	32	29	−3	−3
Netherlands	39	43	37	+4	−6
United States	31	25	18	−6	−7

[a] No change in union share of total employment due to fall in agriculture employment.

Source: U.S. Department of Labor, Bureau of Labor Statistics, Office of Productivity and Technology, Division of Foreign Labor Statistics and Trade, July 1986. Center for Labour Economics OECD Data Set, updated with respective country statistical abstracts.

unionization. Even read cautiously, however, they clearly contradict the notion that the decline in union density in the United States is part of a general collapse of unions in the developed world; in most countries, union density increased in the 1970s and stabilized or declined modestly in the 1980s at levels above those in earlier years. In Japan, where union density fell in both decades, the rate of decline was half as large as in the United States, and less significant in terms of the entire work force because the proportion of Japanese working as nonagricultural wage and salary workers increased, maintaining the union share of total employment. In Canada, where many of the same companies and unions operate as in the United States and where living standards and culture (exclusive of Quebec) are similar, union density went from roughly the same to twice the U.S. level from 1970 to 1985. A persuasive explanation of the decline in union density in the United States should also explain why density did not decline in Canada in the same time period. A third country to which I would like to draw attention is the United Kingdom, where density fell in the 1980s, probably at rates above those shown in the table (the union figures come from the unions, who exaggerate membership in a period of decline); where organization in the strongest union structure, the closed shop (which requires workers to be union), fell sharply; where an increasing proportion of private manufacturing firms operated nonunion; where employment grew more rapidly in low-union density than in high-union density plants; and where proportions covered by collective contracts also fell noticeably.[4] Because union density increased in the 1970s in the United Kingdom, however, these changes have not produced anything like an American-style drop through 1985.

Overall, the patterns of change shown in Exhibit 4 highlight the fact that de-unionization is largely, if not exclusively, a U.S. development.[5] This fact has important implications. First, it disallows any broad explanation of the decline, say, that "unions have become obsolete in the modern 'post-industrial' market economy" (whatever those grandiose terms mean). Second, it constitutes powerful evidence against the structuralist argument that changes in the composition of the work force and jobs from (traditionally union) male blue collar labor to (traditionally nonunion) female, white collar, and service labor underlie the drop. If "post-industrial" or structural changes inexorably reduce unionization, density would have fallen in Canada and other developed capitalist countries, all of which have experienced essentially the same structural changes as the United States, as well as in the United States. Finally, the divergent country trends also cast doubt on any general macroeconomic explanation of U.S. de-unionization: it is hard to argue that the economic problems that followed the oil shock, the inflation of the 1970s, or ensuing deflation caused unions to decline in the United States when the same forces did not reduce union density elsewhere. Put differently, the international data direct attention away from worldwide economic developments to the particulars of American labor relations.

How American Workers Organize

"No, no! The adventures first," said the Gryphon in an impatient tone: "explanations take such a dreadful time."

As reported by Lewis Carroll

Before evaluating proposed explanations for the change in union density in the United States, it is important to understand the way in which workers are organized.

---

[4] Millward and Stevens (1986) report a rise in the percentage of manufacturing establishments operating nonunion from 18 percent to 29 percent (pp. 58–59) and a drop in the percentage with a closed shop from 30 percent to 18 percent (p. 103).

[5] This is not to say that unions throughout the West do not face serious problems in adjusting to the changing economic climate. They do. The difference between the United States and most other countries is that only in the United States have these problems taken the form of massive de-unionization of the private sector.

In contrast to some countries such as the United Kingdom where unionization is left largely to private parties, governments in the United States regulate extensively the process of union representation in both private and public sectors.

The National Labor-Management Relations Act (NLRA) of 1935, amended in 1947 (Taft-Hartley Act) and afterwards, establishes secret ballot representation elections for private sector workers to choose whether or not to become union and regulates the labor relations conduct of employers and unions in several ways. The Act forbids employers to discriminate against workers for union activity and forbids unions to engage in certain tactics deemed unfair as well, such as secondary boycotts.[6] Under the National Labor Relations Board process, unionization currently takes the following form: a union petitions for an election by producing the signatures of 30 percent or more (in practice a union will petition for an election only if two-thirds of the employees are willing to request an election). The NLRB decides the set of workers eligible to vote, and, subject to legal objections and appeals that can delay the actual election for several months, supervises the voting. If a majority votes for the union the NLRB will certify it as the collective bargaining representative of the work force, which obligates the employer to bargain, though not to come to agreement. To resolve impasses, workers can strike while employers can lock out employees or hire strikebreakers. The NLRB also supervises secret-ballot "decertification" elections for unionized workers who wish to reconsider their status.

The key to understanding the process is to recognize that although only workers vote in a representation election, management plays an active role in most campaigns, generally trying to convince workers to reject unions. One study of some 200 organizing campaigns in 1982–83, for example, estimated that the direct supervisors of workers played a "sizeable" or "extreme" role in two-third of the elections, while abstaining in just 6 percent (AFL-CIO, Appendix, p. 37). Many managements employ labor-management consulting experts ("union-busters") to convince workers to vote against the union. Determining union status through an adversarial election framework which accords management a major role is a distinctly American process which has contributed, I will argue, to the observed decline of private sector union density.

The fact that each state has its own public sector labor law means that there is no uniform method by which workers unionize in the public sector. Indeed, while some states enacted NLRA-type laws that create secret ballot election procedures for workers to decide union status and that require employers to negotiate with the unions, often mandating compulsory arbitration to resolve disputes, other states outlaw unions or collective bargaining.[7] Because state laws often mimic the NLRA, however, the situation is not as chaotic as might first appear. A reasonable generalization is that in states with comprehensive public sector labor laws, the de jure process of unionization in the public sector is much like that in the private sector, with workers deciding union status through government-sponsored voting procedures. As the method for unionization is comparable between the public and private sectors, an explanation of divergent sectoral trends must explain why comparable processes produce different outcomes in the two sectors.

---

[6] That is, situations where a union tries to pressure an employer by placing economic pressure on a third party, often a customer or supplier.

[7] The changes in law are reviewed in Freeman (1986) and in Valletta and Freeman (forthcoming) in greater detail. The NBER has a computer file on regulations across states, departments, and time, which is available to researchers on request.

Finally, U.S. labor law establishes representation elections for workers to choose union status but does not outlaw private agreements between workers and management regarding unionization. Unions and employers can agree that a given workplace will be organized (as GM, Toyota, and UAW have done for the Fremont, California plant) with an understanding that if the workers or another union seek an election, management will remain neutral or encourage workers to choose the preferred union. Unions can also sign up workers and demand recognition, threatening a strike should management refuse. In some sectors, moreover, employers are union because they hire from union hiring halls. While non-NLRB modes of organization are likely to become more important in the future, representation elections have been the predominant mode of organization in the past thirty years, making the outcome of the election process critical to union growth in the period under study.

Relating New Organization to Density

The discussion thus far has dealt with the process by which workers choose to unionize and hence with the flow of new union members. As our ultimate concern is with changes in union density it is essential to relate the flows to the "stock" of union members. The following identity, which makes total union membership a function of the changes in existing membership and of "investment" through organizing activity, relates the flows to the stocks:

$$\text{UNION}(t) = \text{UNION}(t-1) - r\text{UNION}(t-1) + \text{NEW}(t).$$

> UNION is the number of union members (persons covered by collective bargaining contracts) in a given year t.
>
> $-r$ is the rate of change in membership due to changes in employment in organized establishments. It may be negative or positive, though the "natural" process by which some union plants close yearly (while new plants are born nonunion) suggests that $-r$ will be negative, implying a natural depreciation in union membership, even in a growing economy.
>
> NEW is the number of new members obtained through organization of new workplaces or lost through decertification of unions at existing workplaces in the period from $t-1$ to t. For reasons of data availability NEW is best thought of as net members gained through NLRB elections, though in principal one would like to include the result of all new organizing activity.

Dividing both sides of the identity by the number of employees in year t to measure union density (UDENS), and manipulating, one obtains

$$\begin{aligned}\text{UDENS}(t) &= [1/(1+g)][(1-r)\text{UDENS}(t-1) + \text{PCTNEW}(t)]\\ &\approx (1-r-g)\text{UDENS}(t-1) + (1-g)\text{PCTNEW}(t)\end{aligned}$$

PCTNEW is the ratio of workers organized to the work force in $t - 1$; g is the rate of growth of total employment; and $-(r + g)$ is the net depreciation or appreciation of union density.

The steady state level of union density implied by this difference equation is UDENS = PCTNEW/$(r + g)$, which shows that the permanent level of union density changes whenever $r + g$ or PCTNEW change; that is, whenever economic conditions cause union plants to contract/expand relative to the growth of total employment or when the rate of union organization changes over time. For example, if unions organize 1 percent of the work force per annum, as they did in the 1950s, and suffer a net depreciation of density of 4 percent, the union share of employment would stabilize at 25 percent. If, alternatively, new organization fell to .7 percent of the work force, as it did in the 1960s, and $r + g$ remained the same, the union share would drop to 17.5 percent in the long run.

How important are changes in new organization versus changes in the rate of depreciation of existing union membership in the recent collapse of private sector union density? Crude calculations applying our difference equation to private sector union membership and number of workers won through NLRB elections show a rising trend in the depreciation rate in recent years: estimated compound annual rates of $r + g$ for the 1950s of 4.0 percent; for the 1960s, 3.4 percent; for the 1970s, 4.7 percent; and for 1980–85, 6.1 percent.[8] Two factors are likely to underlie the increase in $r + g$ in the 1970s and 1980s: the concentration of unionism in slow-growth areas and potential increased death of union plants due to the rising union wage premiums of the seventies. The crudeness of the data underlying the estimates of $r + g$ makes one leery, however, of taking estimated changes too seriously. What should be taken seriously is the negative value of $-(r + g)$, which implies that unions, like the Red Queen in *Through the Looking Glass* for whom "it takes all the running you can do, to keep in the same place," must organize large numbers of workers each year to maintain private sector density.

That unions have failed to do this through the NLRB electoral process is evident from the statistics summarized in Exhibit 5. Columns 1 and 2 measure the success of unions in organizing workers through NLRB representation elections in selected years from 1950 to 1983: column 1 gives the absolute number of workers won by unions in representation elections while column 2 gives the number of workers won relative to the work force. These data show that union success in representation elections has declined to the point where virtually no workers are organized through NLRB

---

[8] These estimates are based on the private sector union membership numbers of Troy and Sheflin for 1950, 1960, 1970, 1980, and a projected 1985 figure based on his 1983 (updated using CPS figures on the percentage change in membership from 1983 to 1985); estimates of private nonagricultural wage and salary workers exclusive of private household workers from the BLS; and the number of members won by unions in NLRB elections in 1951–1960, 1961–1970, 1971–1980, 1981–1985 (with 1984 and 1985 assumed to be the same as 1983). I calculated the compound annual rate of depreciation in union membership by comparing actual membership in, say 1980, with membership in 1970 plus the number of workers won from 1971–1980, and then subtracted the compound annual rate of growth of the relevant employment.

Exhibit 5

Organizing success of unions through NLRB elections

	Workers won (in thousands)	*Workers won / nonagr. emp.*	*% of elig. in union wins*	*Elig. workers / nonagr. emp.*
1950	754	2.0	84	2.4
1955	343	1.0	73	1.4
1960	286	.7	59	1.2
1965	316	.7	61	1.1
1970	301	.6	52	1.2
1975	204	.4	38	1.1
1980	173	.2	37	.5
1983	91	.1	43	.3

Source: NLRB Annual Reports.

procedures. Whereas in the early 1950s unions organized 1 percent to 2 percent of the work force via government-sponsored elections, in the 1960s they organized about .7 percent; in the 1970s, about .5 percent; and in 1983, just .1 percent: 91,000 workers in a work force of some 90 million! Even these figures, moreover, understate union inability to gain members through NLRB elections for a union electoral victory does not guarantee that workers will obtain a collective bargaining contract. In recent years, in fact, workers voting to unionize failed to gain a contract approximately a third of the time (Weiler, 1985), and thus worked under management-determined rather than collectively bargained conditions. The implication is that in 1983 perhaps only 60,000 workers gained union representation through the electoral process. When account is taken of the 25,000 or so workers who chose to decertify unions in elections, it is apparent that the legally established mode of organizing labor in the private sector of the United States has run dry for trade unions.

Columns 3 and 4 of Exhibit 5 decompose the number of workers won by unions in representation elections relative to the work force into: a measure of the union success rate—the number of workers in elections won by unions relative to the number of workers eligible to vote (column 3); and a measure of the extent of electoral activity—the number of workers eligible to vote in elections relative to the work force. The figures in column 3 show that one factor in the decline in workers organized through representation elections has been a fall in the proportion of workers in elections won by unions (which reflects declines in union win rates and in the average size of union victories).[9] The figures in column 4 show, however, an even greater proportionate drop in the extent of electoral activity: in 1960 some 6000 NLRB elections covered 1.2 percent of the work force; in 1983 4400 elections covered a bare

---

[9] One can use various identities to decompose the decline in members won/employment (Freeman, 1985).

.3 percent of the work force. Had unions won all of these elections they would not have gained enough members to increase density.

Turning to the public sector, there are good reasons to believe that both terms in our equation for union density—the net depreciation of existing density $-(r + g)$ and success in organizing new workers—played a role in the speed with which unionization and/or collective bargaining representation grew in the 1970s and 1980s. First, because few government departments, union or nonunion, go out of business, the loss of density due to closure of union workplaces (reflected in r) will be small. Second, the relatively moderate growth of public sector employment into new areas implies that existing unions do not have to appeal to workers outside their traditional jurisdictions to maintain their share of employment, which also should make $r + g$ smaller in the public than in the private sector. Third, organizing workers into unions has undoubtedly been much easier in the public sector. In part this is because many workers have been historically organized into employee associations, which were readily transformed into unions without massive organizing campaign. More importantly, however, organization is easier in the public sector because politicians and public sector managers do not contest unionization to the extent that private employers do, for reasons to be laid out shortly. Finally, state laws that require employers to negotiate and mandate impasse procedures guarantee that workers who vote union ultimately receive union representation. Hence, in the public sector, once workers choose union representation, they get what they choose.

With this and the information in the preceding section as background, let us now turn to proposed explanations for the changed union density.

Potential Causes of Change

"It's rather hard to understand... Somehow it seems to fill my head with ideas—only I don't know exactly what they are! However, somebody killed something: that's clear, at any rate."

Alice, as reported by Lewis Carroll

Researchers have proposed a wide variety of factors to explain the changes in unionization in the United States. Among the hypothesized causes for the decline in private sector density are: structural shifts in the composition of the work force and mix of jobs; changes in public attitudes toward unions reflected in opinion polls; increased governmental regulation of the labor market substituting for union protection; "positive labor relations" by nonunion firms; the performance of unions in representing workers and in allocating resources to organizing drives; antiunion policies of the Reagan administration evinced in the air traffic controllers' strike and selection of members to the NLRB; antiunion campaigns by managements. Among the factors said to cause the spurt in public sector unionism are: extension of NLRA type laws to public sector workers in many states; extension of some union rights to federal workers by executive order; pent-up demand for unions by workers whose organization level was exceptionally low by U.S. and world standards.

By examining the impact of factors on the changes in the private and public sectors at once, and requiring that the factors have a consistent effect, I eliminate some hypotheses and highlight the role of others.

Structural Changes

Because the proportion of the work force in traditionally nonunion occupations (white collar), demographic groups (females, college graduates), industries (service and trade) and regions (the South) has increased rapidly, various analysts have explored the possibility that the changing structure of the work force underlies the drop in union density. The tool for this exploration is a fixed coefficient model which decomposes the workforce into a number of groups with varying degrees of unionization at some base time, and explores the impact of changes in the relative size of the groups on aggregate union density under the assumption that the density of each group is fixed at its base level.

Fixed coefficient analyses covering from the 1960s through 1980 attribute 50 percent to 70 percent of the decline in private sector union density to compositional factors with the increase in white collar employment having a particularly sizeable depressant effect on density. While these calculations would appear to go far toward explaining the decline in unionization, I believe the structuralist analysis is misleading and should be rejected. There are three reasons for rejecting it.

First, the structuralist hypothesis is inconsistent with the rise in union density in other countries (notably Canada) which had structural changes in the work force similar to those in the United States. Second, surveys of worker desire for unionism show that structural changes cannot explain the decline in union success in NLRB elections, since the groups whose proportion of the work force increased (such as women and young workers) have as great or greater desire to unionize as do white prime-age male workers (Freeman and Medoff, 1984, p. 228). Third, I reject the structuralist hypothesis because it assumes that the union share of workers in a sector should remain fixed over time. That assumption is inconsistent with the history of union growth, which is one of expansion into nonunion areas, as occurred in the public sector in the period under study. The claim that public sector workers organized because of pent-up demand for unionization, indeed, implies that unionism "naturally" grows in new sectors over time. Fixed coefficient calculations sidestep the key issue in the decline in U.S. unionization, which is why unions failed to organize historically nonunion workers in the private sector while doing so in the public sector and in other countries.

Public Opinion of Unions

The decrease in favorable attitudes toward unions shown in public opinion polls offers another possible explanation for the decline of unionism in the private sector (Lipset, 1986). While public opinion data should not be dismissed out of hand, I do not find this hypothesis persuasive. The timing of the change in opinions and union density are at best weakly related, with public support of unions as measured by the Gallup Poll climbing from 62 percent in 1949 to 75 percent in 1953, falling to 64 percent in 1962, rising to 70 percent in 1965, falling to 55 percent in 1981, and rebounding slightly to 58 percent in 1985. Public favorableness toward unions was 59 percent in 1973 and 58 percent in 1985, but union density fell sharply during that time. More damaging to this thesis, perhaps, are responses to the behaviorally more meaningful question of whether workers would vote for a union at their work place, which show no decline in worker desire for unions: in both 1977 and 1984 about a third of nonunion workers said they wanted unions at their workplace.[10] Finally, a public opinion explanation of union decline is inconsistent with unionization of the public sector, where public opinion could be expected to be especially important.

Substitutes for Union Protection

Workers join unions for protection against unfair treatment by management, including low wages. This motivation suggests that the development of substitute modes of protection in the form of welfare state interventions in markets (Neumann and Rissman, 1984) or of better personnel practices by management might reduce worker desire for union representation. If government activities substitute for union protection, one would expect: unionism to decline most in countries with the greatest welfare state and restrictions on management, in American states with extensive protective labor legislation, and in periods of declining governmental interventions. Further, unions should appeal least to workers who enjoy special legal protections, such as blacks, women, and public sector employees protected by civil service rules. Changes in union density by country, by states, and over time, and the desire for unions by blacks, women, and public sector workers are uniformly inconsistent with these implications (Freeman, 1986b). Hence, I reject the view that governmental regulations and growth of the welfare state underlie the decline in density.

The possibility that "positive labor relations" practices adopted by many major companies (but too expensive for smaller ones)—paying union level wages and instituting union-style personnel practices such as seniority, job bidding and posting, grievance systems (Foulkes, 1980)—has reduced worker desires for unions at these workplaces finds better support in the data. Studies of organizing drives show that companies with good personnel practices are more successful than others in defeating unions (AFL-CIO, 1984; Kochan, McKersie, and Chalykoff, 1986), although the largest effect of such practices must be in deterring drives in the first place. The increased wage premium associated with size of establishment in the 1970s (Brown and Medoff, 1986), which could be expected to make unionization less attractive to those workers, is, moreover, consistent with the declining success of unions in NLRB elections involving large establishments.[11] Still, I doubt that positive labor relations can explain much of the overall decline in union density. Large employers account for

---

[10] The figures for 1977 are from the Quality of Employment Survey. Those for 1985 are from the Harris Poll.

[11] In the 1950s, unions had a better record in large units than in smaller ones, but this pattern of success was reversed in the 1970s. The small number of elections in large units in recent years makes the 1980s figures spotty. For data on union win rates by size of unit, see Goldfield (1987).

a decreasing share of jobs in the U.S. and unions could increase density greatly without organizing the IBMs of the world.

Government Industrial Relations Policies

Government industrial relations policies have been cited as causing unions to decline in the private sector and to expand in the public sector. For the private sector, some blame (credit) the Reagan Administration's operation of the NLRB and destruction of the Professional Air Traffic Controllers Organization (PATCO) as inducing an antiunion climate in the business community. The fact that union density began falling in the private sector before the Reagan Administration and that collective bargaining coverage has increased in the public sector during the 1980s would seem to rule out this hypothesis, though it is possible that administration actions may have contributed to the acceleration of the decline in the 1980s.

The argument that state policies toward unionism have played a major role in the expansion of collective bargaining in the public sector has, on the other hand, considerable research support. Numerous studies have found that public sector unionization and collective bargaining contracts are more likely in states with favorable labor laws than in other states (Freeman, 1986). Other studies—perhaps more persuasive—have shown that unionism has spurted in states following passage of favorable laws (Ichniowski, forthcoming; Saltzman, 1985, forthcoming). Even within a city, departments for which state laws are more favorable toward collective bargaining end up with contracts more frequently than "brother" departments operating under less favorable state law (Freeman and Valletta, forthcoming). At the minimum, we have an empirical regularity: public sector collective bargaining coverage and union density increase markedly in the presence of laws favorable to collective bargaining. This finding raises, however, two further questions: why comparable laws induce different outcomes between the public and private sectors, with which I will deal in detail shortly; and why many states enacted laws favorable to unions while "labor law reform" failed at the national level, which I believe is due in part to the simple fact that a minority of senators, representing states unfavorably disposed to unionism, can stop national legislation but cannot, of course, undo legislation in states that are favorably disposed to unionism.

Union Performance

Union performance may have contributed to the decline in union density in three ways. First, unions may have represented members poorly, discouraging nonunion workers from organizing. This hypothesis has little empirical support; it is inconsistent with union success in obtaining wage and benefit increases exceeding those of other workers in the 1970s and with opinion poll data showing union members to be reasonably satisfied with their unions (Kochan, 1979). Second, unions may have failed to allocate sufficient resources to organizing activity in the 1960s and 1970s. This claim has some a priori validity, as organizing expenditures deflated by wages (organizing is labor-intensive) have failed to keep pace with the growth of the increasing nonunion labor force and have been concentrated in sectors in which unions were already strong rather than in new and growing industries (Paula Voos, 1983; 1984). Because decisions to allocate resources to organizing depend in part on the organizing environment and the perceived benefits and costs of organizing campaigns, however, I am leery of attributing a large independent role to union organizing efforts. Union failure to embark on major organizing campaigns may simply reflect rational responses to the low expected rates of success. Third, and paradoxically from the perspective of analyses that treat unionization as a worker decision, the success of unions in raising union wage premiums in the 1970s may have contributed to the decline in density by raising the cost of unionization to firms and thereby intensifying the antiunion activity of management—to which I turn next.

Aggressive Antiunion Management

Given that the NLRB electoral process allows management to influence results, and that management is likely to prefer to operate nonunion, it is logical to seek an explanation for declining union success in the behavior of management. Did antiunion activity by management increase in the private sector in the period studied? How have public sector managers behaved?

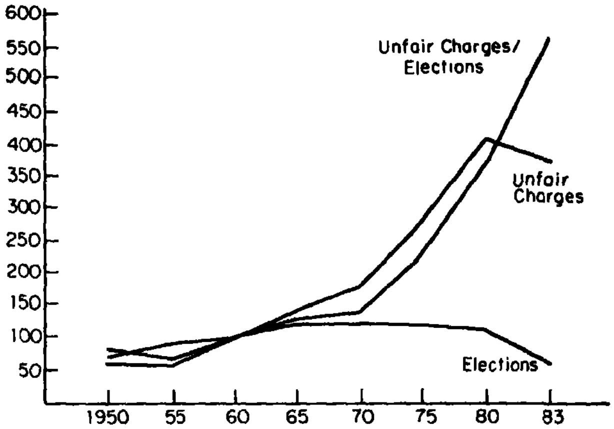
Exhibit 6 shows graphically that one indicator of management antiunion activity—"unfair" labor practices under the law—has skyrocketed. While the NLRB data in the exhibit measure charges of unfair activity rather than actions found illegal in court (raising the possibility that unions may simply be filing more unfair charges), the proportion of charges upheld has been roughly constant at 30 to 40 percent over time (Weiler, 1983), and NLRB statistics on numbers of workers discharged for union activity and ordered reinstated by the NLRB and courts show increases comparable to those in the figure. As for legal management opposition, a Conference Board survey reveals that 45 percent of firms in their Personnel Practices Forum had "operating union free" as a labor policy goal in 1983 compared to 31 percent in 1977, indicating that even over a short period management opposition to unionization has grown substantially (Kochan, McKersie and Chalykoff, 1986).

In the public sector, by contrast, there has been no outburst of antiunion activity by management. Charges of unfair labor practices concern interpretation of state bargaining laws—whether a particular topic is subject to collective bargaining or is a management prerogative—not opposition to the existence of unionism per se. Public sector managers rarely hire union-preventing firms to discourage organization by their workers. Because of changes in state laws, moreover, the trend in the public sector is toward less rather than more management opposition to collective bargaining.

On the basis of these facts, I argue that the antiunion management offensive in the private sector is the key to de-unionization of the United States, and that its absence from the public sector explains the successful organization of public employees. Moreover, I claim that the differential behavior of management in the public and private sectors is explicable by the incentives facing them.

Exhibit 6

Indices of charges of unfair labor practices against management and NLRB elections (1960 = 100)



Source: NLRB Annual Reports.

Management Offensive, Union Wage Premiums, and Cost of Opposition

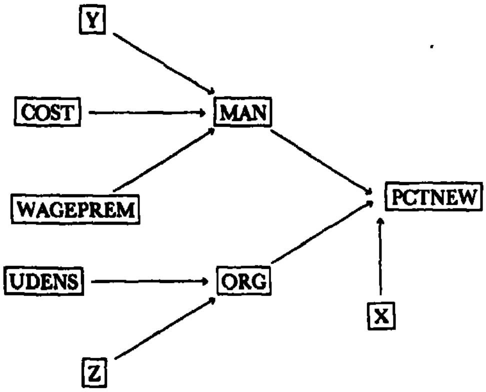
My proposed explanation can be most easily represented with a simple schematic model (Exhibit 7), which diverges from many models of union organization by stressing the role of management in unionization and the endogeneity of both management opposition and union organizing. I postulate three basic relations.

The first relation is a production function for organizing success. It relates the number of workers newly organized relative to the labor force to the resources devoted by management to opposing unions, the resources devoted by unions to organizing, and a vector of all other factors that might influence the outcome.

The second relation links the resources management devotes to opposing unionization to economic factors likely to affect the profitability of such activity. One such factor is the union wage premium, which is assumed to reduce profits and increase management opposition. Another is the cost of opposing unionism in an NLRB election, which will depend on the "technology" of battling unions at workplaces (which has changed greatly in the past fifteen years due to the advent of labor-management consultants who specialize in training supervisors to pressure workers to oppose unions and in running antiunion propaganda campaigns) and on the legal penalty for committing unfair labor practices. In the private sector the legal penalty is minuscule: the law requires that management reinstate workers unjustly fired for union activity, pay the workers back pay less whatever income they earned in the

Exhibit 7

Determinants of union organizing activity and management opposition and their effect on unionization of new workers



The first relation

Determination of organizing success: PCTNEW = f(MAN, ORG, X)
where PCTNEW = number of workers organized/labor force;
MAN = resources devoted by management to opposing unions;
ORG = resources devoted by unions to organizing;
X = other factors that influence outcomes

The second relation

Management opposition: MAN = g(WAGEPREM, COST, ORG, Y)
where WAGEPREM = union wage premium;
COST = cost of opposing union in NLRB elections;
Y = other relevant factors, largely relating to product market
factors such as deregulation of industries, etc., which determine effect of unionism on profitability.

The third relation

Union and worker organizing effort: ORG = h(WAGEPREM, MAN, UDENS, Z)
where UDENS = union density at the beginning of the period;
Z = other relevant factors, largely relating to labor market.

period, and post a notice that the firm will not engage in such illegal activity again. However, reinstating workers often occurs only after the representation election and many choose not to return. Posting a notice about past illegal activity on the part of the firm often has the effect of warning workers how far management is willing to go to defeat unions rather than convincing them that management will forego such tactics in the future. Finally, the management decision is also assumed to depend on the resources the union devotes to organizing, and on other unspecified factors, largely relating to product market conditions, that can be expected to make anticipated union wage premiums more or less expensive in terms of lost profits.

The third relationship in my analysis links union (and worker) organizing activity to: the wage premium, presumed to raise the attractiveness of unions to workers and thus increase organizing activity; the existing density of union workers, with a lower density assumed to reduce organizing activity because the cost of organizing each nonunion worker is higher to existing union members when there are proportionately fewer unionists among whom to spread the cost; the amount of resources management devotes to opposing unionism; and a catch-all vector of other relevant factors, largely relating to labor market conditions.

One important feature of this model is that the union wage premium affects the decisions of both the management and union (workers). In the management equation, a higher premium induces more antiunion activity and thus reduces union organizing success. In the union equation, a higher premium induces additional organizing activity and thus raises union organizing success. Given opposing tendencies, is the wage premium likely to be positively or negatively correlated with actual outcomes? To the extent that unions extract rents from firms through monopoly wage increases, the money loss to firms will exceed the transfer to workers (due to the "triangle" inefficiency), giving management a potential incentive to spend more resources to prevent unionization than unions/workers spend to organize (Freeman, 1986a). If expenditures by the two sides have equal effects on outcomes, higher union wage premium will reduce organizing success.

A second important feature of the model is the inclusion of management and union activity in the equations determining the behavior of the other side. This highlights the interactive nature of the organizing struggle—which can be viewed as a two person (three person, if one distinguishes workers at an organizing site from the union) game—and thus directs attention to the potential payoff from game-theoretic analyses of union and management strategic behavior that goes beyond the scope of this essay.

A third aspect of the analysis that deserves attention is the inclusion of union density in the equation for union organizing activity, for it raises the possibility that a fall in union density will produce a cumulative decline, as the increased cost of organizing induces unions to lower organizing activity, while, conversely, rises in union density will have the opposite cumulative effect.

How might one use the model embedded in these relationships to explain the decline in private sector union density?

The most reasonable hypothesis, given the evidence in the previous section, is that management opposition to unionism induced by changes in the cost of unionization and the union wage premium reduced union organizing success, with cumulative effects on union density and organizing effort. For this explanation to be valid, it is necessary that management activity has an important impact on organizing success; and that this activity responds to economic incentives.

With respect to the effectiveness of management opposition, studies by a diverse set of researchers ranging from management groups to the General Accounting Office to the AFL-CIO to academics find that management activity reduces union success in NLRB representation elections and organizing drives (see Exhibit 8). The sole

Exhibit 8

Summary of studies of effects of management activity on NLRB representation election results

Study and sample	*Measurement of management activity*	*Does activity have effect?*
1. National Industrial conference Board, 140 union organising drives of white collar workers, 1966–67	Amount of communication by management	Yes
2. AFL-CIO, 495 NLRB election, 1966–67	Amount of opposition by management	Yes
3. Prosten (1978), analysis of probability of union win in 130,701 elections in 1962–77	Amount of time delay between election and petition	Yes
4. Lawler (1984), 155 NLRB elections, 1974–78	Company hires consultant	Yes
5. Drotning (1967), 41 elections ordered void and rerun by NLRB	Amount of communication by management	Yes
6. Roomkin-Block (1981), 45155 union representation cases, 1971–77	Delay between petition and election	Yes
7. Seeber and Cooke (1983), proportion of workers voting for union representation by state, 1970–78	Employers object to election district	Yes
8. US General Accounting Office (1982), analysis of 8(a)(3) illegal firings or other discrimination for union involvement in 368 representation elections	Employer committed unfair labor practice	Yes
9. Aspin (1966) study the 71 NLRB elections in which reinstatements were ordered.	Employer fired worker for union activity	Yes, unless reinstated before elect
10. Getman, Goldberg, and Herman (1976) analysis of 1293 workers in 31 elections in 1972–73	Campaign tactics employer.	Not stat-significant
11. Dickens (1983) study of 966 workers in 31 elections in 1972–73 (using data set in #10)	Legal and illegal campaign tactics by employer	Yes
12. Catler (1978) study of 817 NLRB elections	Unfair labor practices and delay	Yes
13. Kochan, McKersie and Chalykoff (1986), 225 firms,	Employer emphasizes union avoidance strategy	Yes

exception (Getman, Goldberg, and Herman, 1976) has been the subject of considerable controversy, with Dickens convincingly reversing their conclusion in a re-analysis of the underlying data. Even absent this, however, the preponderance of the evidence is that *the extent of management opposition substantially determines the outcomes of organizing campaigns*, as posited. As management devotes considerable resources to opposing unionization, and presumably acts rationally, this makes intuitive sense.

Because the studies in Exhibit 8 focus on representation elections/organizing drives, rather than on the workers gained relative to the work force (PCTNEW) that

Exhibit 9

Estimates of the effect of management unfair labor practices on percentage of the nonagricultural workforce newly organized in the NLRB elections

	Impact of 10% increase in unfair practices / election on proportion of workers newly organized in NLRB elect.	*Estimated % decline in organized workforce in NLRB election due to increased management unfair labor practices*
Comparison of union success		
Across states, 1950–78	−2.5	28
Within states over time, 1950–78	−3.4	38
Across industries, 1965–80	−3.5	36
Within industries over time, 1965–80	−6.2	62
In US over time, 1950–80	−4.6	51

Source: Lines 1 and 2 from Freeman and Medoff, 1984, p. 238.
Lines 3 to 5 from Freeman, 1986a.

enters the union density equation, and do not provide estimates of the extent to which the downward trend in union success can be attributed to management opposition, I have estimated the impact of the one indicator of management opposition for which time series data exists—unfair labor practices—on workers won by unions as a share of nonagricultural employment. My analysis uses pooled state (industry) time series data, with time dummies included to control for omitted trend and cyclical factors and state/industry dummies to control for omitted state/industry effects, and diverse control variables. The results, summarized in Exhibit 9, show that unfair labor practices reduce the number of those unionized relative to the work force and that the trend in unfair labor practices can account for roughly half of the observed drop in that outcome measure. To the extent that unfair practices substitute for lawful management antiunion activity, this estimate, while sizeable, understates the full impact of management's fight against unions. Alternatively, however, if the growth of unfair practices is positively correlated with the growth of legal management opposition, my estimates overstate the impact of illegal management activity on the decline in union success, though they may accurately measure the impact of management opposition, in total.

The role of changes in the relative cost of union deterrence in simulating the management offensive has not been the subject of extensive research, making conclusions here more speculative. Time series calculations show that some of the rise in unfair practices is explicable by the rise in union wage premiums in the 1970s, but the series is short and the measures are crude (Freeman, 1986a). Estimates of the relation between decline of union density and union premiums at a one-digit industry level indicate that where the premium has risen most, union declines have been greatest

(Linneman and Wachter, 1986), which may reflect reduced organizing success or a greater death rate for union plants in those sectors. As for other factors likely to raise the cost of unionism to firms, it is highly plausible that such product market developments as the growth of foreign competition and deregulation have made existing union wages more expensive in terms of lost profits. As increased nonunion competition and management policies against unionized firms in construction indicates, however, new product market competition from any source, not simply trade or deregulation, can significantly reduce union density. Finally, with respect to the determination of union organizing efforts, the major work here—by Paula Voos—suggests that one can treat organizing behavior as rational decision-making.

Is this analysis consistent with the increase in public sector density?

There are three basic facts to explain regarding the spread of collective bargaining to the public sector: the geographic variation in unionization associated with different labor laws; the spurt in unionization that followed passage of the laws; and the different behavior of public and private sector managers under comparable laws. If the benefits to management of operating with a union relative to the cost of opposing unions are lower in the public than in the private sector, and those costs have decreased over time, we could account for these facts in the same framework used to explain certain developments in the private sector: in terms of the incentives and options for management to oppose unions.

In fact, the incentives for management to oppose unions do appear to be lower in the public sector. First, public sector workers constitute an especially active political group able to punish or reward the politicians who are their employers at the ballot box, even though they are only a small proportion of voters in most areas. Second, the cost of illegal opposition is likely to be greater for public than for private officials, as public officials who break laws are likely to face possible removal from office. Third, unions can help public sector employers increase budgets through lobbying for additional public expenditures, creating a greater jointness of interest than in the private sector. Fourth, wage premiums tend to be smaller in the public sector (see Lewis, forthcoming; Freeman, 1986). Put crudely, management opposition to unions can gain profits in the private sector; in the public sector, it can cost votes. Given this, and the fact that states passed laws favorable to unionization in the 1960s and 1970s that increased the cost of management opposition in the public sector, the spurt of unions there is consistent with the model.

Finally, is this analysis consistent with the differential experience of Canada? While private sector managers are likely to have similar profit incentives to oppose unions in Canada as in the United States (Anderson and Gunderson (1982) show there is no evidence that union wage effects are smaller in Canada than in the United States), the Canadian system for organizing workers differs from the American system by giving management less option to express opposition. Canadian labor boards rely largely on "card checks" in which unions sign up workers to determine representation at a work place rather than on adversarial elections and impose harsher penalties on managements that break the laws (Weiler, 1983; Meltz, 1985), with the result that Canada has not experienced anything like the massive outburst of management unfair labor practices that characterizes U.S. labor relations. The difference in institutional procedures highlights the fact that the decline in private sector unionism in the U.S. required two factors: increased incentive for management to oppose unions, and the opportunity to turn that incentive into action.

Conclusion

As the reader will undoubtedly have observed, not all the pieces for a complete explanation of the decline in private sector union density and rise in public sector density in the United States are in place. But the available evidence seems consistent with an explanation of private sector decline that stresses increased management opposition to union organization, motivated in part by profit-seeking behavior, and augmented by trade union responses; and an explanation of growth in the public sector that stresses reduced management opposition due to passage of comprehensive collective bargaining laws and vote-seeking behavior.

As for the likely impact of the change in unionization on the performance of the economy, two points are worth attention. The first is that the increasingly public sector locus of American unionism is likely to produce different union effects and modes of operation than have been found for the traditionally private sector union movement. Public sector unions have, for example, smaller wage effects and strike rates than private sector unions and appear to increase rather than reduce employment in union activities (see Freeman and Ichniowski, forthcoming). Second, without making any judgment of whether a less unionized private sector will perform better or not, one thing should be perfectly clear: analysts who have attributed national economic problems ranging from unemployment to wage inflation to low productivity to unions will have to find a new culprit to blame: unless there is a remarkable renaissance in unionism, critics won't have unions to kick around any more.

References

American Federation of Labor—Congress of Industrial Organizations, Department of Field Services, *AFL-CIO Organizing Survey*, April, 1984, Washington, D.C.

Anderson, John, and Morley Gunderson, *Union Management Relations in Canada*, Addison, Wesley, 1982.

Aspin, Leslie, *A Study of Reinstatement Under the National Labor Relations Act*, Ph.D. dissertation, MIT, 1966.

Brown, Charles and James Medoff, "The Employer Size Wage Effect," mimeo, Harvard University, 1986.

Catler, Susan, "Labor Union Representation Elections: What Determines Who Wins?" B.A. thesis, Harvard University, 1978.

Dickens, William, "The Effect of Company Campaigns on Certification Elections," *Industrial and Labor Relations Review*, 1983, 36, pp. 560–75.

Dickens, William, and Jon Leonard,

"Accounting for the Decline in Union Membership, 1950–1980," *Industrial and Labor Relations Review*, April 1985, *38*(3), pp. 323–334.

Drotning, John, "NLRB Remedies for Election Misconduct: An Analysis of Election Outcomes and their Determinants," *Journal of Business*, April 1967, *40*(2), pp. 137–48.

Foulkes, Fred, *Personnel Policies in Large Nonunion Companies*. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Freeman, Richard, "The Effect of the Union Wage Differential on Management Opposition and Union Organizing Success," *American Economic Review*, May 1986a, pp. 92–96.

Freeman, Richard, "Unionism Comes to the Public Sector," *Journal of Economic Literature*, Vol. XXIV, March 1986b, pp. 41–86.

Freeman, Richard, "Why Are Unions Fairing Poorly in NLRB Representation Elections?" in Thomas A. Kichan, ed., *Challenges and Choices Facing American Labor*. Cambridge, MA: MIT, 1985, pp. 45–64.

Freeman, Richard, and B. Ichniowski, "When Public Sector Workers Unionise: Summary of Findings." In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Freeman, Richard, and James Medoff, *What Do Unions Do?* New York: Basic Books, 1984.

Freeman, Richard, and Rob Valletta, "The Effects of Public Sector Labor Laws on Unionisation, Wages, and Employment." In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Freeman, Richard, B. Ichniowski, and J. Zax, "Measuring Collective Organization in the Public Sector." In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Getman, Jules, Stephen Goldberg, and Jeanne Herman, *Union Representation Elections: Law and Reality*. New York: Russell Sage Foundation, 1976.

Goldfield, M., *The Decline of Organized Labor*, University of Chicago Press, 1987.

Ichniowski, C., "Public Sector Union Growth and Bargaining Laws." In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Kochan, Thomas, "How American Workers View Labor Unions," *Monthly Labor Review*, 1979 *102*(4), pp. 23–31.

Kochan, Thomas, Robert McKersie, and John Chalykoff, "The Effects of Corporate Strategy and Workplace Innovations on Union Representation," *Industrial and Labor Relations Review*, July 1986, *39*(4), pp. 487–501.

Kramer, Leo, *Labor's Paradox–the American Federation of State, Council, and Municipal Employees*, 1962.

Lawlor, John, "The Influence of Management Consultants on the Outcome of Union Certification Elections," *Industrial and Labor Relations Review*, 1984, *38*, pp. 38–51.

Linneman, Peter and M. Wachter, "Rising Union Premiums and the Declining Boundaries Among Noncompeting Groups," *American Economic Review*, May 1986, *76*, 103–108.

Lipset, Seymour M., "North American Labor Movements: A Comparative Perspective" in Lipset, ed., *Unions in Transition*. San Francisco: Institute for Contemporary Studies, 1986.

Lewis, H. Gregg, "Union/Nonunion Wage Gaps in the Public Sector." In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Meltz, Noah, "Labor Movements in Canada and the United States: Are They Really That Different?" in Thomas A. Kochan, ed., *Challenges and Choices Facing American* (Cambridge, MA; MIT, 1985).

Millward, Neil and Mark Stevens, *British Workplace Industrial Relations 1980–84*. England: Gower Publishing Co. Ltd., 1986.

National Labor Relations Board, Annual Report (Washington, DC; USGPO).

Neumann, George, and Ellen Rissman, "Where Have All the Union Members Gone?" *Journal of Labor Economics*, April 1984, *2*(2), pp. 175–192.

Prosten, Richard, "The Longest Season Union Organizing in the Last Decade," Proceedings of the 31st Annual Meeting of the Industrial Relations Research Association, Madison, Wisconsin, 1978, pp. 240–49.

Roomkin, Myron, and Richard Block, "Case Processing Time and Outcome of Elections: Some Empirical Evidence," *University of Illinois Law Review*, 1981, *VI*, pp. 75–97.

Saltzman, G.M., "Bargaining Laws and Cause and Consequences of the Growth of Teacher's Unions," *Industrial and Labor Relations Review*, 1985, *38*, 335–57.

Saltzman, G.M., "The Effects of Recent Bargaining Laws in Ohio and Illinois," In Freeman, R.B. and C. Ichniowski, eds., *When Public Sector Workers Unionise*. NBER, University of Chicago Press, forthcoming.

Seeber, R., and W. Cooke, "The Decline in Union Success in NLRB Representation Elections," *Industrial Relations*, Winter 1983, *22*(1).

Troy, Leo, and Neil Sheflin, *Union Sourcebook*.

New Jersey: Industrial Relations Data and Information Services, 1985.

U.S. Department of Labor, Bureau of Labor Statistics, *Directory of National Unions and Employee Associations*, various editions.

U.S. General Accounting Office, *Concerns Regarding Impact of Employee Charges Against Employers for Unfair Labor Practices*. Washington, D.C., GAO-HRD 82-80, June 21, 1982.

Voos, Paula, "Union Organizing: Costs and Benefits," *Industrial and Labor Relations Review*, July 1983, 576–91.

Voos, Paula, "Trends in Union Organizing Expenditures, 1953–1977," *Industrial and Labor Relations Review*, October 1984, *38*, pp. 52–63.

Weiler, P., · "Promises to Keep: Securing Workers' Rights to Self-Organisation under the NLRA," *Harvard Law Review*, June 1983, *96*(8), pp. 1769–1827.

Weiler, P., "Striking A New Balance: Freedom of Contract and Prospects for Union Representation," *Harvard Law Review*, 1985, *98* pp. 351–420.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 89–110

The Rise and Fall of Unions: The Public Sector and the Private

Melvin W. Reder

As of the late 1980s, American unionism as a whole has been in decline for about thirty years. This decline has been concentrated in the unions serving workers in the private sector and contrasts sharply with the growth of public sector unions during much of the same period. This paper attempts to explain these divergent trends, but takes their existence as given. By decline (or rise) of unionism I refer to changes in the percentage (relative number) of union adherents among a specified group of workers. Though I believe that movements in the relative number of union adherents have run parallel to broader, but more vague, movements of union "influence," the reader must draw such inferences for himself as best he can.

A union adherent is assumed to be both a dues payer and covered by a collective bargaining contract; possible differences between these two concepts of union affiliation are unimportant for this paper. Individuals are assumed always to act as if they had compared the costs and benefits of union adherence and chosen adherence if the benefits were greater, and otherwise if the opposite applied. As used here, costs and benefits are limited to hourly compensation (wages and fringes), employment opportunities and working conditions. While ideological preferences undoubtedly influence a small minority of workers, it is assumed that such individuals are too few to affect the trends analysed.[1]

The above assumptions obscure important interdependencies among the decisions of individual workers and those of an employee and his employer. For example, if an

---

[1] I do not deny either that changes in attitudes toward the relative desirability of working under union conditions may have occurred from time to time, or that such changes might be of interest. However, I assume that the dominant factors affecting movements in the quantity of union adherents are changes in the benefits and costs of unionization, particularly hourly compensation and employment opportunities.

■ *Melvin W. Reder is Isidore Brown and Gladys J. Brown Professor of Urban and Labor Economics, Emeritus, Graduate School of Business of the University of Chicago, Chicago, Illinois.*

employer is antiunion, any given level of benefits that the union might obtain for an individual will come at a higher cost in terms of striking and/or employee disfavor (should he be a union activist) than if the employer were otherwise disposed.

Also, to an individual, the cost of adherence will be greatly affected by the choices of his fellow workers. Deviation from the majority in a matter as important as union adherence is likely to entail very unpleasant personal relations and unsatisfactory working conditions. Thus, most individuals either conform to the majority view on union adherence or change employers. (I shall bypass the important and difficult problems involved in explaining how consensus is reached or how, once reached, it is altered; such problems are beyond the scope of this paper.)[2]

This paper has a single purpose: to explain why unionism (interpreted as percentage of union adherents among eligible employees) has been declining in the private sector for the past 30 to 35 years, but, for most of this period, has been increasing in the public sector. To provide this explanation, I offer what is formally an ad hoc "theory." It does not derive from a broadly applicable theory of organizational affiliation having a variety of testable implications, but from a rather narrow theory whose observational implications consist of little more than that trends in percentage of union adherents during various time periods have specified directions. These time periods are associated with (allegedly) different levels of cost and benefit of union adherence.

For the private sector, the paper tries to account for differences in the trends of union adherence during five "epochs:" before 1917; 1917 to 1920; 1920 to 1933; 1933 to the early 1950s; since the early 1950s. In the public sector two epochs are distinguished: before and after 1960. Finally, I distinguish between the public and private sectors since 1960. While the paper does comment, in passing, on movements within epochs (that is, year-to-year movements or cyclical variations), these remarks are extraneous to the main theme. It is (almost) as though the year-to-year movements of union adherence *within epochs* were governed by "laws" that differed from the law governing differences across epochs. Obviously, it would be better if there were a single unified theory that accounted for all aspects of the phenomena, but for the time being there is none.

To be sure, there have been several attempts to estimate time series regressions that purported to explain movements of the numbers of union adherents with various regressors serving as explanatory variables. (These include Ashenfelter and Pencavel, 1969; Adams and Krislov, 1974; Bain and Elsheikh, 1976; and Sheflin, Troy and Koeller, 1981.) However, none of these attempts has gained general acceptance and, with or without their assistance, I do not feel confident of my understanding of the interrelated forces that govern the year-to-year movements in the percentage of employees who are union adherents. Consequently, I cannot properly specify a regression equation (or multi-equation model) that would account for such movements.

---

[2] To the individual worker the employer's attitude toward unionism is a datum, but the attitude is not independent of what he (the employer) perceives as the prevailing view among his employees. This is because the cost (in strikes and reduced productivity) of opposing unionism when a majority of the company's employees are pro-union might be so high as to make it advantageous for an employer to abandon the opposition he might otherwise display.

What follows is an attempt to write a "stylized history" of the phenomena in question. Stylized history focuses on certain salient aspects of phenomena while ignoring others, explaining the behavior of what is salient by means of a theory which is (for the purpose of the account) treated as a maintained hypothesis. Its proper field of application is where poverty of data and/or deficiencies of specification make formal econometric methods ineffective.

The Private Sector

Before 1945

In 1919–1920, union membership was at a temporary peak of roughly 5 million, from which it fell, more or less steadily, until it reached a trough of a little less than 3 million around the beginning of 1933. From there it increased to about 9 million at the time of our entry into World War II, and to 17 million or so in the early 1950s (Troy, 1962, pp. 1–2). Although, strictly speaking, this paper refers to events that began in the mid-1940s, the forces operative during the last four decades are not very different from those at work in earlier periods, so that it is useful to start in the years preceding World War I.

I assume that, despite fluctuations, the hourly compensation of union adherents (in the U.S. economy) has always exceeded that of *fully comparable* non-adherents.[3] For most workers the rationale for non-adherence is assumed to be cost (of adherence), especially reduced employment prospects. A major cause of such reduced employment prospects was employer opposition to unionism; among American workers prior to the mid-1930s employer opposition was always a major deterrent to union adherence. This opposition was frequently expressed in refusal to hire known union workers, or to negotiate with their unions, and extended even to compiling and circulating blacklists of known unionists.

Hostility to unionism was also expressed by attempting to obtain permanent replacements for striking workers, so that a striker gambled not only with his current wages, but with his future employment prospects as well. The net effect of these potential expressions of employer hostility was to create a risk of reduced employment opportunities for union adherents that, for many workers, outweighed the potential gain in rate of compensation.

---

[3] The difference in hourly compensation of a union adherent from that of a fully comparable non-adherent (that is, an individual identical in every respect except union adherence) is similar to that of the union-induced wage gap defined by Lewis (1986, p. 15). It is essential to recognize that, as used here, the benefits of unionization (like higher hourly compensation) may be more than offset by an associated reduction in employment prospects.

It is reasonable to suppose that the salience of the threat of reduced employment opportunities varied inversely with the scarcity of such opportunities, which suggests that the perceived cost of union adherence would also have varied inversely with the level of unemployment.[4] So interpreted, the threat of reduced employment opportunities rationalizes the sharp rise in union membership that occurred during the labor scarce period from 1917 to 1920, and the subsequent reduction (from about 5 million to around 4 million) during the deep depression of 1920–21. Further, it accounts for the decline in membership from 1929 to 1933 during the Great Depression (from 3.4 million to 3 million) and for the increase (from 9 to 13 million) during World War II.[5] (All figures presented here are taken from Troy, 1962, Tables 1 and 2.)

Stated briefly, the argument is as follows: to a potential union adherent, other things equal, the perceived cost of union adherence varies with the state of the labor market, being lower when the level of unemployment is low, and conversely. The reason for this is that in a tight labor market—low unemployment—a dismissed worker or replaced striker faces low search costs in obtaining another job. Conversely, in this same situation, an employer faces relatively high search costs in obtaining a new worker. In addition, the high correlation between low unemployment and strong product demand[6] makes employers anxious to avoid loss of current output and accordingly willing to accommodate worker demands, including demands for acceptance of unionization.[7] Thus, in tight labor markets, workers face relatively low costs of quitting or striking, and employers face relatively high costs of resisting worker demands. The argument applies in reverse in conditions of high unemployment.

---

[4] For the period prior to 1945, I assume that the temporal pattern of union adherence among private sector workers was identical with that of private and public combined, because the number of public sector unionists was so small.

[5] This analysis would be consistent with the often remarked pro-cyclical variation in number (and relative percentage) of members that has characterized American unionism since its inception (see Bain and Elsheikh, 1976, Chapter 2).

[6] Two remarks are in order: (1) During both World Wars, labor market tightness was even greater than what would be suggested by the low unemployment percentages taken in isolation. This is because of the open ended demand for military equipment at cost plus prices, which had the effect of creating excess demand for all grades of labor (at going wage rates) in plants producing war goods. In addition, the armed services sharply reduced civilian labor supply and, during World War I, the cessation of European immigration further reduced labor supply. (2) During both World Wars, as part of an effort to prevent strikes, the federal government actively intervened to encourage unionism. During World War I, this effort involved limitation of dismissal of unionists (Perlman and Taft, 1935, Chapter XXXII) and in World War II it involved strong encouragement of compulsory unionism (Seidman, 1953, Chapter 6). Both of these factors reinforced the effect of full employment in reducing the cost of union adherence. Their absence during the Vietman war might account for the failure of that war to trigger an upsurge in union adherence.

[7] The assertions of this paragraph are offered as description, not as analysis. Why employers do not raise rates of compensation sufficiently to avoid tight labor markets—and lower them enough to avoid high unemployment—is a continuing subject of unresolved theoretical debate. The above argument involves commitment to no particular answer to this question; it merely notes the recurrence of such states and adduces their presence to explaining concurrent movements of union adherence.

The argument of this paragraph assumes that some workers with special attributes (like the highly skilled) or employed in special situations (like those protected by a "union label" product market) are unionized in all states of the labor market, but that others can get union jobs only when the labor market is tight. By implication, this means that the latter type of worker prefers to adhere and, in a tight labor market, does so. However, he will not accept the employment loss necessary to remain a union adherent in a non-tight labor market.

Here, as always, unicausal explanations are suspect. A critical reader will ask about the role of the incentive to union adherence, especially about the effect of variations in the union wage gap. Certainly this gap has not been constant through time (Johnson, 1984; Freeman and Medoff, 1984, Ch. 3; Lewis, 1986, Ch. 9) and, intuitively, one would expect that its movements would have had some effect on union adherence.

Nevertheless, I abstract from—ignore—this variable for two reasons: first, from the viewpoint of potential union adherents, other things equal, the attraction of unionism should increase with the union wage gap. But from the viewpoint of employers, the cost of unionization, *and the incentive to resist it*, should also increase with the union wage gap. Failing a model in which these effects can be separately estimated and used to compute the net effect, there is no reason—formal or intuitive—to suppose the effect of this variable (on union adherence) to have one sign rather than the other. Second, to treat the union wage gap as exogenous is intuitively unappealing; to make it endogenous in an estimable model is beyond the present state of the art. Accordingly, and with full recognition of the potential error of omission, I abstract from the effect of the union wage gap on union adherence. "Let him who can do better, proceed."

Obviously, a tight labor market story of union membership behavior will not account for the history of 1933 to 1941. To explain this period special factors must be considered as follows: from 1932 through 1940, legislation (Norris-LaGuardia Act, National Industrial Recovery Act and Wagner Act) and the generally pro-union attitude of the federal government reduced the cost of union adherence and greatly increased the percentage of adherents in the labor force.[8] During the late 1930s, this increase in unionization continued despite high and (in 1937–38) rising levels of the unemployment percentage which, taken in isolation, would suggest that the cost of adherence had been rising, and the percentage of adherents (accordingly) declining. That the reverse happened is due to the aforementioned political climate and legislation, especially the effect of the National Labor Relations Board (established by the Wagner Act) in limiting the ability of antiunion employers to dismiss or otherwise punish union adherents.

---

[8] The Norris-LaGuardia Act of 1932 drastically reduced the power of the federal courts to issue injunctions in labor disputes. The predominant use of such injunctions had been to impede union organizing. It also made "yellow-dog" contracts (contracts in which workers agreed not to join a union as a condition of employment) unenforceable in federal courts.

The National Industrial Recovery Act of 1933 (NIRA) provided for the establishment of industry organizations whose function would be the promulgation of codes of "fair competition." The Act stated that such organizations should include (among other things) representatives of labor and one section (7a) specified that each industry code should provide for collective bargaining.

The Wagner Act of 1935 replaced NIRA which had been declared unconstitutional in 1935 and which, in any case, had been unenforceable—especially as regards collective bargaining. Wagner stated (in effect) that the policy of the federal government was to promote collective bargaining, in pursuit of which it established a National Labor Relations Board (NLRB) whose function it was to prevent employer interference with the declared right of employees to organize and choose bargaining representatives. Especially during its first few years (1935–41), the Board was extremely aggressive and effective in carrying out its mandate under the law.

Working symbiotically with the National Labor Relations Board (NLRB), the newly formed Congress of Industrial Organization (CIO) unions and (a bit later) the revivified affiliates of the American Federation of Labor (AFL) vigorously proselytized potential members.[9] It is impossible, and for our purposes unnecessary, to disentangle the separate contributions of legislation and (drastically) increased union organizing activity during the period from 1935 to 1941; it is sufficient that their joint effect was to reduce the cost of adherence and increase the relative number of union adherents.

It is instructive to consider a (possible) difference between two sub-periods, 1933–37 and 1937–41. During the latter period the constitutionality of the Wagner Act had already been confirmed, which reassured potential adherents that their right to unionize would be protected. But during 1933–37, union organizing operated under the continuing threat of job loss from hostile employers, with only uncertain protection provided by one provision of the National Industrial Recovery Act and by the generally benevolent attitude of the federal government. Nevertheless, during this period (as well as later) union membership rose sharply, which raises a question as to the salience of the Wagner Act and NLRB for explaining union growth during 1937–41. Would not union organizing activity and/or a favorable shift of worker attitudes toward unionism be sufficient to explain the rise in adherents?

As a speculative *obiter dictum*, I offer the following answer: as with other major legislative changes altering the scope of legally permissible employment practices (like fair employment legislation), statutory enactment followed several years of intense controversy and attempts to pressure reluctant employers to conform "voluntarily" to the spirit of legislation proposed, but not yet enacted. This social pressure aroused hopes for governmental assistance to embattled unionists, and encouraged belief that the number of union adherents was already so large that employers could not easily replace them in the event of a strike.

To a prospective union adherent such hopes would have caused a reduction in the perceived cost of adherence. But such hopes would have been brittle, and sensitive to news concerning their prospects for realization. Legislative defeat of Wagner, or an adverse Supreme Court decision as to its constitutionality, might well have destroyed the fragile tissue of confidence on which union adherence (then) rested. Conversely, employer opposition to unionization was conditioned on hopes of a subsidence of its current popularity; legislative or judicial defeats of Wagner would probably have stiffened their (still) strong resistance.

---

[9] Implicitly, I assume that organizing activity is endogenously determined by the opportunities for success, and is therefore highly correlated with the realized growth and decline of union membership. Also I assume that organizing activity has the effect of reducing the cost of adherence; that is, organizing activity reduces the cost of acquiring information for prospective union adherents and/or builds confidence that the (necessarily uncertain) effort to unionize a bargaining unit will succeed. So characterized, union organizing activity is analogous to advertising expenditure by a firm selling a service.

In short, I interpret rapid union growth during 1933–41, despite very high unemployment, as the joint result of: (1) decline in the cost of union adherence that resulted, initially (1933–37), from the perception of the government's pro-union attitude and the anticipation of effective pro-union legislation, and later (1937–41) from realization of that hope; (2) the organizing efforts of the CIO and AFL and, during the early 1930s, of grass-root organizers. Arguably, there may also have been an increase in pro-union attitudes during this period. However, this story does not require such a shift, and I do not assume that it occurred.

During World War II, "over-full" employment might be sufficient to account for the sharp increase in the unionized percentage of the labor force. However, in addition to the state of the labor market, the federal government actively promoted unionization by encouraging quasi-compulsory unionization in plants devoted to military production (Seidman, 1953, Chapter 6).

Sources of Union Growth and Decline: Overview

The preceding paragraphs account for certain outstanding features of the profile of union membership prior to 1945. But they do not provide a handle for grasping the less dramatic year-to-year movements over long periods of comparative quiescence. To provide such a handle I distinguish between the adherence proclivity of workers currently employed in unionized situations (for example, bargaining units) from that of workers otherwise situated.[10] Where union membership is associated with job security and/or preference in rehiring, the net benefit of adherence often becomes appreciable even where the individual beneficiary is indifferent or even hostile to unionism.

As a result, workers previously employed in such a unionized situation have a strong incentive to remain unionized, and will (in effect) collude to impose unionization on those newly hired. Most often this is accomplished by making (permanent) employment conditional upon union membership. Thus, where unionism is already established, membership is compelled to keep in step with growing employment.

By contrast, in nonunion situations, growth in number of adherents is more difficult to achieve. From the viewpoint of a union organizer, it is necessary to persuade nonmembers of the advantages of adherence; that is, to convince them that the benefits outweigh the costs and steer them away from "free riding." Simultaneously, employers must be convinced that increased resistance is not worth the entailed cost.[11] As a result, the impact of employment growth on number (and percentage) of union adherents is much larger if the growth occurs in industries and/or operations that are already unionized. Thus an important determinant of the growth rate of union adherence is the direction and strength of the correlation (across employment situations) of employment growth with previous unionization.[12]

---

[10] "Unionized situations" include new operations (especially in construction) where the employer agrees to give preference in hiring to union workers before any workers have been hired.

[11] The interaction between persuasion of workers (that employer resistance will not make unionization "too costly" or "too unyielding of benefit") and persuasion of the employer (that worker determination is such as to make resistance unprofitable) is difficult to model but, fortunately, unnecessary for the present purpose. However, the reader should be aware of the analytical problems that are skirted in this statement.

When employment in unionized segments of the economy is in decline, unions must replace the incumbents of disappearing jobs with recruits from previously unorganized employment situations. Very often, unions have been unequal to this task. Consider the case of craft unions of skilled workers: technical progress has posed a continuing threat to their monopolies of labor supply by facilitating replacement of their members by less skilled substitutes. In the process many of the (typically small) craft unions were destroyed, while unionization of their less skilled replacements often proved impossible. This process of union demolition was a major and often remarked feature of the history of the AFL into the 1930s (Slichter, 1941, especially Chapters XII and XIII). Its obverse was recurrent complaint (from union protagonists) about the failure of unskilled mass production workers to organize.

A second source of union job erosion has been geographical competition in product markets. Even before the 1930s, union wage scales set in one part of the country were undermined by lower wage rates paid elsewhere to workers of comparable efficiency, resulting in a decline of union employment and membership. Inter-area competition of this kind was prominent in bituminous coal mining, textiles, clothing manufacturing and book and job printing (Slichter, 1941; Millis, 1942, Ch. 3, 5, 8, 9). Such competition had the effect of reducing the employment of union members and, by limiting the union wage gap, of reducing the incentive for new workers to adhere.

Curbing inter-area competition, both in product and labor markets, was a major objective of the National Industrial Recovery Act (1933–35) and of the CIO unions formed in 1935–36. The desire (need) to accomplish this objective gave impetus to the organizing drives of the later 1930s, which were bitterly resisted by nonunion employers bent on maintaining their labor cost advantage over unionized firms. The resulting struggle was temporarily halted during World War II when both tight labor markets and governmental policies favorable to unionization (particularly compulsory membership) discouraged inter-area competition (Seidman, 1953, Ch. 6).

In a nutshell: prior to 1945 American unions "had to run hard to stand still" (Dunlop, 1980, p. 400) and, save for three exceptional periods, they were unsuccessful in the attempt.[13] This theme is not new; it has been developed by a number of writers and, at least in broad outline, has not been seriously disputed. (Among these writers are Davis, 1941; Dunlop 1948, 1980; Reder, 1957; Galenson, 1960; Rezler, 1961; Freeman and Medoff, 1984.) It is my contention that, with a few modifications, the same story applies to the post-1945 period.

---

[12] I expect that the assertions in this paragraph will be generally accepted; moreover, in principle, they are subject to empirical test. However, at present, there is no available data set from which a time series of actual union membership as a percentage of "workers eligible for membership," for unionized and non-unionized situations separately, could be calculated. Hence these assertions must, for the near future, be treated as conjectures. In their interesting paper, Dickens and Leonard (1985) attempt to estimate the net growth of union membership due to "economic causes." While their results are not inconsistent with the general tenor of this paper, there are sufficient differences of concept and purpose as to make detailed comparison both difficult and, for the purpose at hand, irrelevant.

[13] Dunlop (1980) mentions five exceptional periods: two of these are in the 19th century and one is 1951–53. The last of these is (implicitly) dealt with (see below) as part of the post-World War II employment growth in previously unionized firms. Rationalizing the 19th century examples would require a more complicated story than I have presented here.

From 1945 to Circa 1960

From 1945 until 1953, the share of union adherents in the labor force increased, but thereafter the trend was downward. The postwar increase is in sharp contrast to the decline of union membership after World War I. One obvious reason for the difference is the absence of a major depression, like that of 1920–21, after World War II. A second reason is the obstacles to "union-busting" contained in the Wagner Act and continued in Taft-Hartley; these did not exist prior to 1935. A third reason is that, as a result of the extensive unionization of the 1930s and 1940s, for the first time in American history unionization was promoted on a large scale by expansion of employment in unionized situations rather than by "organization of the unorganized."

As applied to unionized manufacturing firms, the relation of union growth to employment increase is obvious. But the relevance of this mechanism extends beyond manufacturing; in at least two important cases, Teamsters and construction, substantial growth of union membership occurred throughout the late 1940s and 1950s without need of persuading individual workers of the advantage of union adherence.

Teamsters. In this case, organization of new workers was "assisted" by refusal of (already) unionized drivers to make deliveries to nonunion establishments or to accept transfers of freight from nonunion (or other union) carriers. To avoid prolonged, costly and (at that time) seemingly hopeless struggles with the union, employers would persuade their workers to join the Teamsters. (For an excellent discussion of Teamster organizing tactics, see James and James, 1965, especially Chapters 7–11.)

The organizing push of the Teamsters during the late 1940s and 1950s may be interpreted as a continuation of work begun in the 1930s that had been temporarily halted (or slackened) because of the wartime ban on strikes.[14] Throughout the 1950s, the Teamsters' economic muscle (that is, their ability to impede delivery) survived the Taft-Hartley provisions against secondary boycotts, and enabled them to continue recruiting members regardless of the disinterest (or possible distaste) of the individuals concerned.[15] However, after 1960, the provisions of the Landrum-Griffin Act (1959) and the cumulative effect of adverse judicial rulings, gradually undermined the ability of the Teamsters to gain recruits through coercion (Levinson, 1980, pp. 135–37).[16]

Construction. Membership growth of the construction unions during the 1950s was partly a matter of increasing employment in closed shop situations not unlike that in heavy manufacturing. But in addition, in many large projects the difficulties of recruiting skilled construction workers for jobs in out-of-the-way locales in the face of boycotts from relevant unions caused contractors (often at the insistence of owners concerned with completion dates) to sign pre-hiring union shop agreements which made adherence a condition of employment (Mills, 1980, pp. 60–71). This was especially the case in the public sector where buyers, particularly the federal government (in projects concerning atomic energy, military construction, and interstate highways), were notorious for being more concerned with completion dates than with price.

---

[14] This is oversimplified, though true in essence. I omit all the complications resulting from internal division within the Teamsters. For details see James and James (1965), especially Chapters 1, 4–6.

[15] However, it should not be overlooked that (possibly) many workers who might have disapproved of Teamster behavior in the abstract cheerfully accepted its fruits, and gave voting support to Hoffa and his successors, so long as the Teamsters could deliver benefits to the members.

[16] The Landrum-Griffin Act of 1959 was primarily concerned with regulating the internal affairs of unions. Inter alia, it provided a "bill of rights" for union members; established financial reporting requirements (to the U.S. Department of Labor) for unions and limited the ability of national unions to dominate the affairs of local unions. It is this last provision that is relevant in the present context.

To this was added the effect of the Davis-Bacon Act at the federal level, and of "little Davis-Bacon Acts" in the various states (Northrup, 1984, pp. 36–8).[17] During the 1950s and 1960s, the effect of these laws was to limit the ability of cost conscious buyers in the public sector to avail themselves of cheaper nonunion labor. This increased the availability of (high-paying) union jobs and induced union adherence among construction workers.

To summarize: during the late 1940s and throughout the 1950s, expansion of employment in previously unionized manufacturing plants, and in situations like freight hauling (Teamsters) and construction where union ability to impede production could be deployed to coerce employers, caused roughly equal increases in the number of union adherents. Until 1953, these sources of union growth, together with new organizing, were sufficiently strong to promote growth in the *percentage* of union workers, as well. But thereafter, the forces making for relative union growth were overbalanced by growth of employment in nonunion situations, and by erosion of union jobs as a result of the resumption of inter-area competition. That is, after 1945, there was a resumption of the historic pattern of low wage nonunion firms in southern and/or small towns undercutting union firms. The resumption may have been assisted by a provision of the Taft-Hartley Act of 1947 that made invalid, in any state proscribing them, collective bargaining agreements requiring union membership as a condition of employment. As of the present, twenty states (mostly in the south and mid-west) have enacted "right-to-work" laws pursuant to this provision.

The intent of these laws is to encourage employers to divert production and jobs to low wage nonunion areas wherever possible. This tends to limit the size of the union wage gap, and thereby to reduce the benefits of unionism to present and prospective union adherents. However, the actual effect of right-to-work legislation on union membership is a matter of ongoing debate. (For citations, see Flanagan, Smith, and Ehrenberg, 1984, pp. 374–75.) It should also be noted that right-to-work provisions appear to have encouraged non-adherence of individual workers in unionized situations.[18]

---

[17] The Davis-Bacon Act of 1931 requires that wages paid construction workers employed on projects in which there is financial involvement of the federal government shall be not less than the "prevailing wage in the area" as determined by the Secretary of Labor. The effect of the legislation has been to require payment of the "area" union wage scale on all federally funded construction. (Very often, "area" has been interpreted so as to make the applicable union scale as high as possible.) Although the Act is still enforced, recent interpretative regulations have somewhat abated its upward impact on labor costs (Northrup, 1984, p. 42–4; Mills, 1980, pp. 89–92.) Thirty-five states have, at one time or another, enacted laws ("little Davis-Bacon" acts) modeled on the federal Act.

Not only has inter-area domestic competition resumed since 1945, but, beginning in the early 1960s and growing steadily in importance, international competition has made inroads into American markets. This foreign competition has not only intensified the effect of domestic interregional competition on wages and prices in unconcentrated and/or partially unionized industries (like textiles and clothing manufacturing), but, even more importantly, has at least partially overcome the wage-price effects of industrywide unionization combined with nationwide oligopoly in the product markets for major durable goods (like automobiles, steel and electronic equipment). As a result, by the mid-1980s demand for protection from "unfair" foreign competition has become a major issue of national economic policy.[19]

Since 1960

In absolute numbers, union membership continued to grow until 1976; but the percentage unionized declined steadily after the mid-1950s.[20] In effect, nonunion employment has grown faster than union employment. Part of the story is the continuing and intensifying erosion of union jobs through inter-area competition, both domestic and international. While imports have been very important, I do not feel able to judge their relative impact (on union adherence) vis-à-vis domestic interregional competition.

Another factor contributing to the decline of union jobs is deregulation of the transportation industries, notably freight hauling and airlines, which had the effect of increasing the cost disadvantage of union operation. Even before deregulation, the high cost of Teamster contracts provided a strong incentive (for employers) to escape unionization (Levinson, 1980, pp. 137–9). But before 1975, regulation of freight rates had made it possible to pass forward the cost of union contracts without fear of competition from nonunion firms. Thereafter, nonunion operation, or concessions from the union, was a condition of survival for a large and growing number of trucking firms.

In construction, after 1960, nonunion competition became increasingly important in large projects as well as small (Mills, 1980, pp. 89–92). The growth of nonunion firms was a lagging response to a long-standing opportunity, with the delay probably the result of tardy recognition that public policy as well as law (Taft-Hartley) would, in the 1950s, have condoned nonunion operation and punished attempts at union harassment of nonunion employers, whether by secondary boycotts or otherwise.[21]

---

[18] Freeman and Medoff (1984), p. 243, report, "In Right-to-Work states, upwards of 20 percent of the workers covered by collective bargaining contracts are not union members, as compared with 10 percent elsewhere in the country." Freeman and Medoff attribute this difference to "free riding." However, some part of this difference may be due to the greater freedom of choice (in right-to-work states) of those individuals whose dislike of unionism is outweighed by the cost of leaving a job in a firm that had become unionized subsequent to the inception of their employment relationship.

[19] For a good brief summary of the relevant phenomena and associated policy issues see Lee Price (1984). For more statistical detail, see W. H. Branson, "Trends in United States International Trade and Investment Since World War II," in Branson (1980), especially pp. 183–236.

[20] According to Troy and Sheflin (1985, Table 3.62, p. 3–14) union membership among private sector workers stood at 17.0 million in 1970 and declined—more or less steadily thereafter—to 16.2 million in 1979 and 13.0 million (preliminary) in 1984.

So far attention has been focused on the (very important) forces causing union attrition. But the forces impeding recruitment of new members are not less important. While there were a number of such forces, for lack of space, I shall concentrate upon one: employer resistance. But before doing so, let me comment briefly upon the others.

Changing Characteristics of the Labor Force. It has been argued that the more rapidly growing categories of the labor force (such as women, southerners, white collar workers) are less favorable to unionization than other workers. While this is partly true, it has been cogently argued (Farber, 1984; Freeman and Medoff, 1984, pp. 228–30) that this factor is, at best, of limited explanatory power, and that its operation depends upon notoriously unstable differences in unionization proclivities across categories of workers.

Declining Effort of Unions to Recruit New Members. It has been argued that in the last two or three decades there has been a decline in union effort at organizing nonunion workers (Freeman and Medoff, 1984, pp. 228–30; Goldsmith, 1983). However, it is difficult to determine whether such a decline has occurred, and its significance if it has.[22] An increase in employer resistance would reduce the effectiveness of any given level of organizing intensity, and would also reduce the optimum fraction of any given level of resources that would be devoted to organizing by a rational union. Accordingly, one cannot infer changes in the organizing effort of unions solely from their expenditures on this purpose.

Government Activity as a Substitute for Union Services. Prior to the 1930s, various individual unions provided their members with unemployment insurance, industrial accident insurance, education, pensions, and so on. But beginning with the New Deal and continuing at least through the 1970s, the public sector became an increasingly important source of such services, and thereby the supplier of a substitute for union-provided benefits. By a standard argument of price theory, this would be expected to cause a reduced demand for union services and, other things equal, a reduction of the percentage of union adherents.

---

[21] In construction, as with the Teamsters, it is plausible to suppose that a large wage gap between union and nonunion wages contributed substantially to the relative growth of nonunion production —particulary after 1970. But to accept this supposition as valid would immediately raise a further question; why did the unions in question set (or permit) the wage gap to become so high? A satisfactory answer would require specification of a model of union wage policy; this is not presently available. Accordingly, I shall not pursue this point, but acknowledge it as an important topic for research. For recent interesting work on the subject, see Flanagan (1987, Ch. 4).

[22] For example, Freeman and Medoff (1984), pp. 228–29, point out that union organizing expenditure per nonunion worker in its jurisdiction (in constant dollars) declined by 30 percent from 1953–74. But while accepting the Freeman-Medoff statement, Paula Voos (1984, pp. 62–63) contends, "More persuasive, however, is the evidence that the total real organizing expenditures of these unions increased and their real expenditures per organizable worker remained relatively constant over the years studied." A further difficulty, that I find perplexing, is that of inferring a shift of some "effort function" from a decline in real resource expenditure. Such a decline may reflect, wholly or in part, a decrease in expected return to effort; in particular, a decrease due to increased employer opposition, and could occur with an unchanged effort function.

In developing this argument, Neumann and Rissman (1984, pp. 186–191) make the interesting point that since the mid-1970s, legislative restrictions (at the state level) on "employment-at-will"[23] have provided a public sector substitute for union job protection against arbitrary dismissal (also see Lipset, 1986; Garbarino, 1984). They present a regression indicating that union adherence has declined more in states where employment-at-will has been restricted than in others. It is also worth noting that civil rights legislation and associated affirmative action programs have constituted a potent substitute for union action in providing job protection for minority workers and females.

Worker Attitudes. It has been contended that since (say) 1950 there has been a decline in pro-union attitudes among workers (Lipset, 1986, Chapter 12). While I do not deny that such a change may have occurred, its significance is difficult to interpret. Evidence from opinion polls or questionnaires (such as Lipset presents) is of uncertain meaning because of the difficulty of distinguishing between disapproval of union behavior per se, and disapproval due to dissatisfaction with union effectiveness in attaining its objectives. The latter cause of disapproval might well result from anticipation of successful employer resistance, rather than from hostility to union methods or objectives.

Employer Resistance. Both in the literature of the subject and in fact, employer resistance has been by far the most important obstacle to union organizing. However, its effectiveness and modus operandi have varied with the political climate and the state of labor law. From 1933 until the passage of the Taft-Hartley Act in 1947, the influence and power of the federal government was exerted to promote the unionization of the American labor force. While the Wagner Act and its vigorous enforcement by the National Labor Relations Board was the most outstanding example of the government's pro-union tropism, it was by no means the only instance. The War Labor Board's encouragement of (modified) union shop provisions and dues checkoff (withholding of union dues from paychecks) contributed greatly to the growth of union membership during World War II. As another example, the Fair Labor Standards Act of 1938 (federal minimum wage law) was intended to—and did—hamper nonunion competition in low wage industries such as textile manufacturing.

This pro-union political climate ended with the enactment of Taft-Hartley. While T-H continued much of the protection of workers' rights to unionize provided by Wagner, it limited the permissible scope of union tactics and expanded that permitted to employers both in contract negotiations and in representation elections. But while the change in the law made by the passage of T-H was important of itself, the signal given by its enactment may have been of even greater significance. Its enactment, which was bitterly opposed by organized labor and occurred despite a veto by President Truman, signalled the end of the era in which embattled unionists could reasonably hope—and opposing employers fear—that the (federal) government would "do something" that would enable them to prevail in a confrontation.

---

[23] "Employment-at-will" is, as the name suggests, the doctrine that in the absence of contractual restrictions, an employer is legally free to terminate an employee at his pleasure.

With the passage of Taft-Hartley, the government became more or less neutral in union-management relations. This change implied an increase in the expected cost to a worker of adhering to a union and a corresponding decrease in the cost of resisting to an employer, thereby reducing the tendency of prospective union recruits to enroll. However, this change did not occur instantaneously, but was reflected in a gradual decline in ability to recruit.

The evidence for the secular decline in ability to recruit in (previously) nonunion situations is the continuing fall in the percentage of union victories in NLRB supervised Representation Elections after 1950.[24] I suggest the following explanation for this decline: prior to Wagner, many firms paid little or no attention to the requisites for avoiding unionization beyond what was required for overt resistance. Taken by surprise at the union-fostering provisions of Wagner, they were sitting ducks for the union organizers of the late 1930s. However, the history of that period educated many of these employers—and eliminated others who proved to be ineducable—as to the need for careful planning with regard to such matters as plant location, compensation policy, personnel selection, managerial style, make or buy decisions, and so on, if unionization was to be avoided.

Reinforced by the exigencies of recruitment in the tight labor markets of the early 1940s and the continuing attrition of the pre-Wagner troglodytes, the personnel and compensation policies of nonunion employers continued to evolve in the late 1940s and 1950s toward positions strategically better situated to resisting unionism.

This process of adjustment may be thought of as involving a choice among three alternative strategies: (1) accept unionism; (2) resist unionism by matching—more or less—union contract terms; and (3) resist unionization while holding hourly wage cost appreciably below union levels. Firms deciding upon strategy (2) offered wages and benefits—especially pensions and health care—at or near union levels, and altered working conditions, disciplinary procedures and job security prospects so as to conform with what unions were currently obtaining for their members.[25] Presumably, the gains to these employers from remaining nonunion consisted of freedom from strikes, greater flexibility in choosing productive techniques (freedom from union work rules) and improved worker attitudes.

---

[24] The success rate declined from 76 percent during 1950–54 to 37 percent during 1975–79 (Dickens and Leonard, 1985, p. 326). Representation elections (authorized under both Wagner and Taft-Hartley) are supervised by NLRB. In such elections workers in a specified bargaining unit vote by secret ballot to determine whether they are to be represented by a union in collective bargaining. More than one union may appear on the ballot, if there is more than one seeking the right to represent workers in a bargaining unit. The union obtaining a majority of the votes cast becomes the exclusive bargaining agent for all workers in the designated bargaining unit.

[25] That is, the behavior of these employers may be interpreted as a manifestation of the "threat effect" of unionization. The union threat effect is discussed by Lewis (1963) and Rosen (1969).

Firms using strategy (3) actively campaigned against union attempts at organization. Successful use of this strategy depended upon selection of type of product, workers to be employed and, most important, location in a "low wage" area. Given a satisfactory solution of this selection problem, a shrewd campaign strategy for the employer was to appeal to (what I conjecture was) the growing sophistication among nonunion workers as to the probable effect *on their own employment prospects* of a successful attempt to bring their wages and conditions up to union levels. That is, I suggest that the declining success rate of unions may have been due, in part, to increased awareness among low wage nonunion workers that their long run employment prospects *in present locations* depended on continuation of sub-union wages and working conditions. To the extent that this was the case, antiunion votes reflected a rational preference for a nonunion status quo, with employer antiunion campaigns serving as a method of spreading information.

This explanation of the growing success of employer resistance to unionization is different from that of some other writers (notably Freeman and Medoff, 1984, pp. 230–39) who emphasize employer unfair labor practices. While their explanation is not incompatible with that offered here—both factors could be operative—it should be noted that the validity of their argument has been challenged by Flanagan (1987, Ch. 4). But for the present purpose, it is not necessary to choose between these alternative explanations of the declining success rate (for unions) in Representation Elections.

Finally, it seems likely that the continuing decline in the unionized percentage of private sector workers has, since 1980, been assisted by the widely perceived hostility of the executive branch and much of the legislature as well. The resounding defeat of the Democratic presidential ticket in 1984, after explicit endorsement by the AFL-CIO, cannot help but have reinforced this effect.

The argument of this paper is directed toward explaining cumulative movements in the "unionized percentage," with turning points left exogenous; that is, turning points are considered to have resulted from major wars and changes in the political-legal climate. Thus, in this story, the decline in the unionized percentage that began in the 1950s continues into the late 1980s with no end in prospect. While I believe that this percentage has a floor somewhere above zero, the only rationalization that I can offer for this belief is not one readily supported by empirical evidence; that is, I believe there will always be employers whose need or short-sightedness impels them to squeeze the quasi-rents of their employees to the point where the latter coalesce in order to resist further exactions.

Public Sector Unionism

The burst of public sector unionization during the 1960s and 1970s was in many respects a reprise of the history of the private sector 30 to 40 years earlier. In both cases, the spurt in union adherence accompanied or followed hard on the heels of changes in labor law at all levels of government that were favorable to collective bargaining. Whether these changes in labor law are considered to be causes of a reduction in the cost of union adherence (as in the private sector) or of an increase in its benefits is largely taxonomic.[26]

At the federal level, the relevant change came via a series of Executive Orders initiated in 1962 by President Kennedy (Executive Order 10988) that extended recognition to unions of federal workers, though limiting the scope of the collective bargaining in which they might engage. At state and local levels a wide variety of statutes were adopted: some merely prescribed that government officials "meet and confer" with union representatives; others specifically required that there be collective bargaining; and still others prescribed compulsory arbitration to terminate bargaining impasses.[27]

In large part, I concur in Freeman's (1986) cogent argument that changes in labor law were the primary cause of the upsurge in public sector unionism during the 1960s and 1970s. However, I would add a caveat to the effect that the result of changes in the law cannot be reliably distinguished from the consequences of the historical process that caused the changes. Changes in labor law do not occur randomly, but result from political struggle that has effects of its own. Thus, it is hard to believe that, if the legal changes occurring in states with a pro-union climate of opinion had instead by some fluke been visited upon strongly antiunion states, the consequences would have been the same.

As with Wagner, the 1960 cohort of legal changes was but one consequence of widespread social ferment. The civil rights movement and resistance to the Vietnam war generated a climate of opinion in which virtually any group had de facto standing to make demands upon constituted authority, and the right to expect dialogue to ensue. In this environment, public sector employee groups could and did make demands, through formal unions and otherwise, backing them with sit-ins and other forms of direct action. Often, legitimizing collective bargaining was simply recognizing a de facto situation, (sometimes) in the hope of inducing a more cooperative bargaining posture in the union. Nevertheless, if demands for improving the formal status of public sector unions could somehow have been ignored, such legislative rejection would surely have discouraged the growth of public sector unionism.

The rapid growth of public sector unions immediately after enactment of favorable legislation suggests that, as in the private sector thirty years before, there had been a pent-up demand for collective bargaining. The analogy with the private sector suggests that this growth would decelerate once the pent-up demand was satisfied, with further growth in membership marching—at most—in step with growth of employment. As we shall see, the record of the later 1970s and the 1980s does not reject this conjecture.

---

[26] That is, in contrast to their private sector counterparts, public sector unionists were rarely subjected to systematic harassment by employers; hence their costs of adherence were mainly dues and initiation fees. As adherence accomplished little until unions were permitted to engage in collective bargaining, one could reasonably argue that the changes in labor law increased the return to a given input of resources (dues, time, and so on) to union activity. But one could contend that, under pre-1960 legal conditions, it cost more to achieve a given result through collective bargaining (which was then an illegal activity) than after the change in law.

[27] For a brief summary of the relevant legislation, see Freeman (1986, pp. 44–9). An earlier but more detailed account is Burton (1979).

For public sector workers, as for private, the incentive for union adherence is the combination of the union wage gap and protection of job opportunities. Many studies of public sector wages find the union wage gap to be smaller than in the private sector. But without arguing the point, I concur in Freeman's (1986, pp. 55–8) reservations to this conclusion. Nevertheless, regardless of its magnitude, the public sector union wage (and fringe) gap is sufficient to attract adherents, which is all that the argument of this paper requires. While it seems likely that (as in the private sector) the public sector union wage gap grew during the first part of the 1970s, I shall not attempt to relate this to the upsurge of public sector unionism.

The attritive forces that eroded the membership base of private sector unions have also operated in the public sector. For two reasons it might be thought that these forces would be weaker in the public sector: first, few, if any, of the services supplied by public sector organizations can be imported from other areas in response to a cost incentive; and second, for a large part of its output, a public sector employer is a monopolist.

However, it is possible to privatize at least part of the public sector, and relatively high public sector costs such as might be caused by a union wage gap would encourage doing so. In addition, in the public sector as in the private, there is substantial opportunity for substituting capital for labor. Still further, differentially high costs of producing locally will be reflected in differential tax rates or service charges undermining local monopolies through the "Tiebout effect" of relocating households or, especially, businesses. (Evidence on this point may be found in an excellent brief discussion in Flanagan, Smith and Ehrenberg, 1984, pp. 475–78.) In short, public sector unionism, like private sector, is subject to erosion from economic forces.

The absolute number of public sector adherents varies with public sector employment, and since the late 1970s there has been very little growth in employment at any level of government.[28] This is in sharp contrast with the rapid growth of public employment in the 1960s and early 1970s and results from well-publicized taxpayers' revolts at state and local levels, and (since 1981) from the campaign of the Reagan administration to reduce taxation and the size of the civilian sector of the federal government.[29]

During this same period, public sector union membership reached a peak of about 6 million in 1976, but declined thereafter and stood at about 5.3 million in 1984 (Troy and Sheflin, 1985, Table 3.62, p. 3–14). Similarly, as Table 1 shows, the unionized percentage also peaked in 1976, at 40 percent, and declined thereafter to about 33 percent in 1984.

[28] Thus, during 1978–86 public sector employment grew from 15.7 to 16.7 million while private sector employment grew from 71.0 to 83.5 million.

[29] From 1981 to 1986 (preliminary data) federal employment grew by 127,000 of which 41,000 were in the resident Armed Forces.

Table 1

Union membership as percent of U.S. labor force: private and public sectors; selected years, 1900–1984

Year	*Total labor force*	*Private Sector*	*Public Sector*
1900	3.0		
1914	6.8		
1917	7.5		
1920	12.0		
1921	11.3		
1922	9.4		
1929	7.0		
1933	5.8		
1937	11.7		
1938	13.4		
1940	14.6		
1945	23.6		
1950	22.1		
1953	26.8		
1960	22.3	31.9	10.8
1962	23.9	31.6	24.3
1970	25.4	29.1	32.0
1971	24 5	28.2	33.0
1972	24.4	27.3	35.4
1973	24.5	26.6	37.0
1974	24.1	26.2	38.0
1975	23.7	26.3	39.6
1976	23.0	25.1	40.2
1977	21.8	23.6	38.1
1978	21.3	22.5	36.7
1979	21.0	22.0	36.4
1980	19.6	20.6	35.1
1981	19.0	19.9	35.4
1982	17.8	19.0	35.1
1983	16.6	17.8	34.1
1984 (p)	16.1	16.6	33.1

Sources: Total labor force: 1900–53, (Bernstein, 1954); 1960–84, (Troy and Sheflin, 1985), Table 3.41; Public Sector.

Note: Since they are from different sources, Column 1 is not strictly comparable to Columns 2 and 3.

Clearly, such a decline would be more than what could be explained by slackened growth in employment. However, these figures, based on number of dues paying members, tell a somewhat different story from other data presented by Freeman (1986, pp. 44–49) which indicates that the unionized percentage more or less leveled off after 1975, but does not suggest a subsequent decline. Accordingly, I shall not speculate on the causes of a possible, but uncertain, decline in the unionized percentage of public sector workers, but merely note that its growth appears to have terminated sometime in the late 1970s.

It is by no means clear that employment growth in the public sector will continue to lag behind the private, but as of 1987 there is no manifest public desire either to increase the federal deficit or the tax bill at any level of government. As a result, further growth of public sector union membership will have to come mainly from organizing in previously nonunion situations. Now that the initial surge after legalization has passed, the only reasons for expecting public sector unions to fare better at new organizing than private unions are the exercise of union political pressure to offset (public) employer efforts to avoid employing unionized workers and the weaker force of inter-area product market competition in the public sector. Whether these two factors will be sufficient to produce relatively greater success at new organizing in the public sector remains to be seen.

Conclusion

It is generally accepted that the outlook for American unionism is uncertain and possibly ominous. Yet its demise is not preordained, nor even likely, and it is quite possible that it may regain the status and workforce coverage that it possessed around the midpoint of the century. But if such a resurgence occurs, it will not be due to the growth of membership in public sector unions. For good or ill, this country is indissolubly wedded to private ownership and operation of the basic means of production, which implies that the bulk of new employment opportunities will continue to be in the private sector. This does not mean that the state's role in economic life will remain at the low level of the middle 1980s—quite the contrary—but that "reintervention" is likely to take the form of partnership or other types of government, business and (possibly) union cooperation.

On the eve of the New Deal, union membership was at low ebb, and the labor movement seemed moribund (Barnett, 1933). Its spectacular revival in the years following was partly due to its role in limiting price competition in product markets; this was initially through the NRA codes of "fair competition," and continued for longer periods in some industries like bituminous coal, clothing manufacturing, textiles and trucking. But during the 1930s, the scope of this role was limited by the fact that durable goods manufacturers were able to maintain oligopolistic price structures without union assistance. However, in the late 1980s, these erstwhile oligopolists are threatened by foreign competition against which they need protection, whatever its source.

To sell such protection to a skeptical public they must stress the (alleged) connection between import exclusion and job protection. In this task, the support of organized labor would be invaluable, if not absolutely essential. Further, by organizing nonunion plants and forcing their unit labor costs to near equality with those of (previously) unionized plants, unions could also provide a brake on inter-area but *intranational* competition.

Still further, since exports have become a significant source of employment and profits, and because any move toward protection against imports would invite export-injuring countermeasures from abroad, producer interests in our own export industries would insist upon agreements with countries adversely affected by our protectionism. Devising such agreements, with the inevitable multiparty balancing of interests within and between competing import and export industries, both here and abroad, would require multiparty negotiations and administration in which unions could make a legitimate demand to participate. In fact, the need for achieving and maintaining consensus among all concerned parties would cause strenuous efforts by other participants to coopt unions.

The reason that a national economic policy to regulate international trade requires consensus is the strong opposition that any (further) departure from free trade would encounter. Such need for consensus would place unions in an excellent position to bargain for pro-union changes in labor law as their reward for participating in the process of planning and administering a "National Economic Policy."[30] Indeed, it was union participation in the formulation and administration of the various industry codes of the National Recovery Administration that helped the unions recover from their 1933 low and paved the way for the Wagner Act of 1935. Even more, it was in exchange for their participation in the war effort (World War II), especially by resisting member demands for strikes and opposing wage controls, that unions obtained the support of the War Labor Board for imposing compulsory union membership and dues checkoffs in plants affecting the war effort.[31]

With these precedents, and demands for a national economic policy to establish "fair trade" widespread, it is not fanciful to imagine a revival of private sector union influence and membership growth associated with a policy of trade intervention designed to protect jobs (and quasi-rents). To suggest this possibility is not to predict that such a drive for protection would be successful and, even less, to applaud it. There is substantial, if quiet, support for free trade and the political struggle is not yet fully joined. Nevertheless, the avowed affinity of American unionism for the cause of protection is clear and, in the event of its triumph, the prospect for benefit to union institutions—though not necessarily to union members—would be quite promising.

■ *This paper has benefited greatly from criticism and suggestions by Robert Flanagan, Edward Lazear, Carl Shapiro and Timothy Taylor. However, the author retains sole responsibility for whatever errors may remain.*

---

[30] A nationally coordinated effort to curb imports while minimizing the consequent restrictions imposed on U.S. exports would require that some form of export quotas be imposed on individual firms (or even plants). Allocating these quotas, coordinating them with gains of jobs and sales from import restrictions and (possibly) with a stimulus to domestic demand, is likely to imply what could reasonably be termed a "National Economic Policy" if not a "National Economic Plan."

[31] The cost to a union for participating in a national economic program is never negligible. In essence, such participation implies acquiescence in actions that impose sacrifices on some of its members for the benefit of others. For a union to do this, it must be protected against the threat that those members who are injured will switch to another union, or abandon unionism altogether.

References

Adams, Arvil V., and Joseph Krislov, "New Union Organizing: A Test of the Ashenfelter-Pencavel Model of Trade Union Growth," *Quarterly Journal of Economics*, May 1974, *88*, 304-11.

Ashenfelter, Orley, and John Pencavel, "American Trade Union Growth, 1900-1960," *Quarterly Journal of Economics*, August 1969, *83*, 434-48.

Bain, George S., and Farouk Elsheikh, *Union Growth and the Business Cycle*. Oxford: Basil Blackwell, 1976.

Barnett, George, "American Trade Unionism and Social Insurance," *American Economic Review*, March 1933, *23*, 1-15.

Branson, William, H., "Trends in United States Investment International Trade and Investment since World War II." In Feldstein, Martin, ed., *The American Economy in Transition*, Chicago: University of Chicago Press, 1980.

Davis, James A., "The Theory of Union Growth," *Quarterly Journal of Economics*, August 1941, *55*, 611-37.

Dickens, William T., and Jonathan S. Leonard, "Accounting for the Decline in Union Membership, 1950-1980," *Industrial and Labor Relations Review*, April 1985, *38*, No. 3., 323-34.

Dunlop, John T., "The Changing Character of Labor Markets." In Feldstein, Martin, ed., *The American Economy in Transition*. Chicago: University of Chicago Press, 1980.

Dunlop, John T., "The Development of Labor Organization: A Theoretical Framework." In Lester, Richard A., and Joseph Shister, eds., *Insights Into Labor Issues*. New York. Macmillan, 1948, pp. 163-93.

Farber, Henry S., "The Extent of Unionization in the United States." In Kochan, Thomas A., ed., *Challenges and Choices Facing American Labor*. Cambridge and London: MIT Press, 1984, pp. 15-43.

Flanagan, Robert J., *Labor Relations and the Litigation Explosion*. Washington, D.C.: The Brookings Institution, 1987.

Flanagan, Robert J., Robert S. Smith, and Ronald G. Ehrenberg, *Labor Economics and Labor Relations*. Glenview, Ill.: Scott, Foresmann and Co., 1984.

Freeman, Richard B., "Unionism Comes to the Public Sector," *Journal of Economic Literature*, March 1986, *24*, 41-86.

Freeman, Richard B., and James L. Medoff, *What Do Unions Do*? New York: Basic Books, 1984.

Galenson, Walter, *The CIO Challenge to the AFL*. Cambridge, Mass.: Harvard University Press. 1960.

Garbarino, Joseph W., "Unionism Without Unions: The New Industrial Relations," *Industrial Relations*, Winter 1984, *23*, 40-51.

Goldsmith, Michael, "Organized Labor in the United States of America: Its Recent Decline and Future Prospects As Reflected in NLRB Union Certification Election Results," Chicago, Department of Political Science, University of Chicago, 1983.

James, Ralph C., and Estelle D. James, *Hoffa and the Teamsters*. Princeton: D. Van Nostrand, 1965.

Johnson, George J., "Changes Over Time in the Union-Nonunion Wage Differential in the United States." In Rosa, Jean-Jacques, ed., *The Economics of Trade Unions: New Directions*. Boston: Kluwer-Nyhoff Publishing, 1984, pp. 3-19.

Levinson, Harold M., "Trucking." In Somers, Gerald G., ed., *Collective Bargaining: Contemporary American Experience*. Madison, WI: Industrial Relations Research Association, 1980, Chapter 3.

Lewis, H. Gregg, *Unionism and Relative Wages in the United States*. Chicago and London: University of Chicago Press, 1963.

Lewis, H. Gregg, *Union Relative Wage Effects*. Chicago and London: University of Chicago Press, 1986.

Lipset, Seymour Martin, ed., *Unions in Transition*. San Francisco: Institute for Contemporary Studies Press, 1986.

Millis, Harry A., ed., *How Collective Bargaining Works*. New York: The Twentieth Century Fund, 1942.

Mills, D. Quinn, "Construction." In Somers, Gerald G., ed., *Collective Bargaining: Contemporary American Experience*. Madison, WI: Industrial Relations Research Association, 1980, Chapter 2.

Neumann, George R., and Ellen R. Rissman, "Where Have All the Union Members Gone?" *Journal of Labor Economics*, April 1984, *2*, 175-92.

Northrup, Herbert R., "Open Shop Construction Revisited," Philadelphia, Industrial Research Unit, The Wharton School, University of Pennsylvania, 1984.

Perlman, Selig, and Philip Taft, *History of Labor in the United States, 1896-1932*. New York: Macmillan, 1935.

Price, Lee, "Growing Problems for American Workers in International Trade." In Kochan, Thomas A., ed., *Challenges and Choices Facing American Labor*. Cambridge and London: MIT Press, 1984, pp. 125-47.

Reder, Melvin W., *Labor in a Growing Economy*. New York. John Wiley, 1957.

Rezler, Julius, *Union Growth Reconsidered: A Cri-*

ical Analysis of Recent Growth Theories. New York: Kossuth Foundation, 1961.

Rosen, Sherwin, "Trade Union Power, Threat Effects and the Extent of Organization," *Review of Economic Studies*, April 1969, *36*, 185–96.

Seidman, Joel, *American Labor from Defense to Reconversion*. Chicago and London: University of Chicago Press, 1953.

Sheflin, Neil, Leo Troy, and C. Timothy Koeller, "Structural Stability in Models of Union Growth," *Quarterly Journal of Economics*, February 1981, *96*, 77–88.

Slichter, Sumner H., *Union Policies and Industrial Management*. Washington: The Brookings Institution, 1941.

Troy, Leo, "Trade Union Membership, 1897–1962," NBER Occasional Paper 92, New York, National Bureau of Economic Research, 1962.

Troy, Leo, and Neil Sheflin, *Union Sourcebook: Membership, Structure, Finance, Directory*. 1st ed. West Orange, N.J.: Industrial Relations Data Information Services, 1985.

Voos, Paula B., "Trends in Union Organizing Expenditures, 1953–1977," *Industrial and Labor Relations Review*, October 1984, *38*, 52–63.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 111–135

In Honor of Sandy Grossman, Winner of the John Bates Clark Medal

David M. Kreps

Sandy Grossman is Professor of Economics at Princeton University, and formerly held positions as Professor of Economics at the University of Chicago, Professor of Economics and Finance at the University of Pennsylvania, Research Associate at the Federal Reserve, and Assistant Professor of Economics and Finance at Stanford University. He is the winner of the John Bates Clark Medal for 1987, given to "that American economist under forty who is adjudged to have made a significant contribution to economic thought and knowledge." I have been invited by the editors to review some of the contributions that Grossman has made to economics.

My review will concentrate on two of the major themes in Grossman's work: models of rational expectations equilibria in situations where there is privately held information; and models of the modern corporation, emphasizing incentives for managers. The selection of these topics was driven primarily by my own interests and expertise, and while these two general topics comprise a sizeable fraction of Grossman's work, they do not exhaust it. The concluding section includes a brief mention of some other themes, as well as some general remarks about the broad impact that Grossman has had on the community of economists.

I had originally intended to close this article with a few stories that illustrate facets of Grossman's, shall we say, strong character. I have certainly had fun collecting stories from many of our mutual colleagues. But Sandy has informed me that he knows a very good libel lawyer. And he is still just in the first blush of youth; it will be far more interesting to sketch out over a longer time period the development of his remarkable personality. The stories I have collected will therefore have to wait the many years until it is time to prepare his Festschrift. Meanwhile, the interested reader may wish to consult the December 1985 issue of *Esquire*, page 388ff.

■ *David M. Kreps is Professor of Economics, Graduate School of Business, Stanford University, Stanford, California.*

Table 1 contains a list of Grossman's published work (as per his curriculum vitae of March 1987). Throughout this review, references to Grossman's papers will be according to the numbers in Table 1, while references to the work of others is by the author (date) system. To avoid overburdening the paper with technical details, I take expositional license and assert a number of things which are subject to qualification or depend, sometimes nontrivially, on unstated assumptions. Hence what follows is meant only to give the flavor of Grossman's research and the issues it addresses.

Table 1

The published work of Sandy Grossman

1. "Rational Expectations and the Economic Modelling of Markets Subject to Uncertainty: A Bayesian Approach," *Journal of Econometrics*, 1975, *3*, 255–72.
2. "Equilibrium under Uncertainty and Bayesian Adaptive Control Theory." In Day, R., and T. Groves, eds., *Adaptive Economic Models*. New York: Academic Press, 1975, pp. 279–307.
3. "On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information," *Journal of Finance*, 1976, *31*, 573–84.
4. "Information and Competitive Price Systems," (with J. E. Stiglitz), *American Economic Review*, 1976, *66*, 246–53.
5. "On Value Maximization and Alternative Objectives of the Firm," (with J. E. Stiglitz), *Journal of Finance*, 1977, *32*, 389–402.
6. "A Characterization of the Optimality of Equilibrium in Incomplete Markets," *Journal of Economic Theory*, 1977, *15*, 1–15.
7. "The Existence of Future Markets, Noisy Rational Expectations and Informational Externalities," *Review of Economic Studies*, 1977, *64*, 431–49.
8. "A Bayesian Approach to the Production of Information and Learning by Doing," (with R. E. Kihlstrom and L. J. Mirman), *Review of Economics Studies*, 1977, *64*, 533–47.
9. "A Theory of Competitive Equilibrium in Stock Market Economies," (with O. Hart), *Econometrica*, 1979, *47*, 293–330.
10. "Further Results on the Informational Efficiency of Competitive Stock Markets," *Journal of Economic Theory*," 1978, *18*, 81–101.
11. "Consumption under Uncertainty," (with D. Levhari and L. Mirman). In Green and Scheinkman, eds., *General Equilibrium, Growth, and Trade*. New York. Academic Press, 1979, pp. 105–24.
12. "Take-Over Bids, the Managerial Theory of the Firm and the Free Rider Problem," (with O. Hart). In Currie, David A., and Will Peters, eds., *Contemporary Economic Analysis*. Vol. 2. London: Croom Helm, 1980, pp. 461–68.
13. "On the Impossibility of Informationally Efficient Markets," (with J. E. Stiglitz), *American Economic Review*, 1980, *70*, 393–408.
14. "On Stockholder Unanimity in Making Production and Financial Decisions," (with J. E. Stiglitz), *Quarterly Journal of Economics*, 1980, *95*, 543–66.
15. "Corporate Financial Structure and Managerial Incentives," (with O. Hart). In McCall, John J., ed., *The Economics of Information and Uncertainty*. Chicago: University of Chicago Press, 1982, pp. 107–40.
16. "Disclosure Laws and Takeover Bids," (with O. Hart), *Journal of Finance*, 1980, *35*, 323–34.
17. "Takeover Bids, The Free-Rider Problem, and the Theory of Corporation," (with O. Hart), *Bell Journal of Economics*, 1980, *11*, 42–64.
18. "Nash Equilibrium and the Industrial Organization of Markets with Large Fixed Costs," *Econometrica*, 1981, *49*, 1149–72.
19. "Implicit Contracts, Moral Hazard and Unemployment," (with O. Hart), *American Economic Review*, 1981, *71*, 301–7.
20. "The Determinants of the Variability of Stock Market Prices," (with R. Shiller), *American Economic Review*, 1981, *71*, 222–8.

Rational Expectations Equilibria

Grossman's thesis and much of his early work concerned the use of rational expectations equilibria to analyze markets in which prices aggregate and transmit private information from agent to agent. It may be helpful, in reviewing his contributions to this area, to set the stage by reviewing the basic concepts involved. (I am cribbing heavily in this recapitulation from [22].)

In many contexts of interest, a decision taken today will turn, at least partly, on prices that will prevail in the future. Agents base their decisions on what they *expect*

Table 1
Continued

21. "The Allocational Role of Takeover Bids in Situations of Asymmetric Information," (with O. Hart), *Journal of Finance*, 1981, *36*, 253-70.
22. "An Introduction to the Theory of Rational Expectations Under Asymmetric Information," *Review of Economic Studies*, 1981, *68*, 541-59.
23. "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, 1981, 461-83.
24. "Savings and Insurance," (with L. Weiss). In Boyer, M., and R. Kihlstrom, eds., *Bayesian Models in Economic Theory*. Lausanne: Elsevier Science Publishers B. V., 1984, pp. 303-311.
25. "Heterogeneous Information and the Theory of the Business Cycle," (with L. Weiss), *Journal of Political Economy*, 1982, *90*, 699-727.
26. "An Analysis of the Principal-Agent Problem," (with O. Hart), *Econometrica*, 1983, *51*, 7-46.
27. "Consumption Correlatedness and Risk Measurement in Economics with Non-Traded Assets, and Heterogeneous Information," (with R. Shiller), *Journal of Financial Economics*, 1982, *10*, 195-210.
28. "Implicit Contracts under Asymmetric Information," (with O. Hart), *Quarterly Journal of Economics*, 1983, *98*, 123-56.
29. "Monetary Non-Neutrality When Prices Are Observables," (with L. Weiss). In Sarnat, Marshall, and Giorgio Szego, eds., *Savings, Investment, and Capital Markets in an Inflationary Economy*. Cambridge, Mass.: Ballinger Publishing Co., 1982, pp. 313-4.
30. "Integration," (with David Evans). In Evans, David, ed., *Breaking Up Bell, Essays on Industrial Organization and Regulation*. New York: North Holland Publishing Co., 1983, pp. 95-126.
31. "A Transactions-Based Model of the Monetary Transmission Mechanism," (with L. Weiss), *American Economic Review*, 1983, *83*, 871-80.
32. "Unemployment with Observable Aggregate Shocks," (with O. Hart and E. Maskin), *Journal of Political Economy*, 1983, *91*, pp. 907-28.
33. "Customer protection in Futures and Securities Markets," (with D. Fischel), *Journal of Futures Markets*, 1984, *4*, 273-95.
34. "The Costs and Benefits of Ownership: A Theory of Vertical Integration," (with O. Hart), *Journal of Political Economy*, 1986, *94*, 691-719.
35. "Economic Costs and Benefits of the Proposed One-Minute Time Bracketing Regulations," (with M. Miller), *Journal of Futures Markets*, 1986, *6*, 141-66.
36. "An Analysis of the Role of 'Insider Trading' on Futures Markets," *Journal of Business*, 1986, *59*, 5129-46.
37. "Vertical Integration and the Distribution of Property Rights." Forthcoming in Razin, A. and E. Sadka, eds., *Economic Policy in Theory and Practice*, 1987.
38. "Sequential Bargaining under Asymmetric Information," (with M. Perry), *Journal of Economic Theory*, 1986, *39*, 120-54.
39. "Perfect Sequential Equilibrium," (with M. Perry), *Journal of Economic Theory*, 1986, *39*, 97-119.
40. "Monetary Dynamics with Proportional Transactions Cost and Fixed Payment Periods." Forthcoming in Barnett, W., and K. Singleton, eds., *New Approaches to Monetary Economics*.

future prices to be. If the future is certain, a point estimate will do, and even when future prices will depend on things which, from today's perspective, are random, in some cases a point estimate is sufficient. But in other cases, the agent will find it necessary to form an assessment of the entire distribution of the future price, or even an assessment of the joint distribution of that price and other variables relevant to the agent's decision. These assessments may reflect past experiences of the agent, and they may be based on reasoning from first principles concerning how the economy is constructed and how it will evolve. This idea is not especially modern; in [22], an example of reasoning from first principles, drawn from Marshall, is given.

When the agent's understanding of the economy is so good that his or her assessment coincides with the prices that will prevail as a function of future contingencies, the agent's expectations (or assessments) are said to be *rational* or to exhibit *perfect foresight*. An equilibrium in which all agents have rational expectations is called a *rational expectations equilibrium*.

Now imagine that, today, different agents in the economy have access to different information which is relevant to predicting what circumstances will prevail in the future. For example, imagine an agent who wishes to forecast the future spot price of wheat in order to decide today how much wheat to place into storage. If this agent obtains (private) information that indicates that the harvest is likely to be disappointingly small, he will anticipate that the distribution of spot prices in the future will be higher than is generally held to be the case. Thus he will store rather more wheat today. This act will withdraw some wheat from current supplies, and (if a number of agents have similar information and act accordingly) the spot price of wheat today will be bid up. Spot prices today are higher because some agents have and act on private information that indicates higher spot prices tomorrow.

Whereas in the simple story of perfect foresight future prices depended on future contingencies (which would presumably be revealed at the time to all interested parties), here it is the current price that is a function of current information which in turn is known to some interested parties and not to others. If the otherwise uninformed parties nonetheless have rational expectations about the relationship between prices and the privately held information, those nominally uninformed parties can use current equilibrium prices (and Bayes' rule) to infer from current prices something about the privately held information. For example, if the current spot price of wheat is relatively high, then an otherwise relatively uninformed party, who lacks, say, information pertaining to the size of the upcoming harvest, will revise upwards his/her assessment of tomorrow's spot price.

Of course, the notion that prices contain and convey information is standard doctrine among economists. Grossman is fond of quoting Hayek on this point, and I will simply follow him in this regard:

> We must look at the price system as... a mechanism for communicating information if we want to understand its real function... The most significant fact about this system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take

Sandy Grossman

> the right action... by a kind of symbol, only the most essential information is passed on... (Hayek, 1945, p. 527)

This idea includes a classic (Walrasian) notion of "information communication" by prices, which should be distinguished from the information communication function of prices in a rational expectations equilibrium. In the Walrasian tradition, prices communicate to agents something about the relative values of goods economy-wide. From the price of a particular good, the agent learns how much it will cost to consume it or what can be gained by producing another unit of it relative to other goods; things that the agent doesn't know *a priori*. But learning that others place a high relative value on the good doesn't change the agent's own private valuation of the good. In the classical tradition, an agent doesn't conclude that wheat is really (personally) tastier than corn just because the price of wheat today is high.

In the context of rational expectations, however, one is learning from prices what other people know, and that information is conceivably useful in formulating one's own tastes. At its simplest, this concerns the direct utility that one will obtain from consumption of a good. For example, if one sees that car X retails for more than car Y, then one may infer that car X is worth more to *oneself*, because one suspects that others have experience with these cars and have come to learn that car X is more reliable, or whatever. But examples extend beyond cases of directly consumed goods to cases where the good is purchased as an asset, to be resold later. Then information in today's price may concern prices (and hence value upon resale) tomorrow. Grossman (in [22], p. 557) makes the following argument in favor of this in the context of the stock market:

> In a Walrasian model, people trade and at the equilibrium price vector have no desire to trade further. Thus, if the next day there is a big change in a price and a person has no change in preferences or income, he will in general

> trade away from his position. Many people hold shares of stock. The stock price varies enormously from day to day. Yet many people do not trade. That is, I come home at the end of the day and learn that the price of a stock which I own has gone up 20%. No other news has impacted on me. If there is no information in the event that the price has gone up 20%, then I would surely sell, since yesterday I was indifferent at the margin and transaction costs are only 2%. I don't sell when I get the news of the price rise because it contains information about the payoff for the stock which leads me to change my preferences for it.

Before moving to Grossman's contributions, I must distinguish between two types of models of rational expectations and raise two fundamental difficulties with the basic concept. First the distinction.

Suppose that an uninformed agent sees a high price for wheat. One reason that the price may be high is that the informed parties have received information which indicates that prices tomorrow will be high. Hence it can be inferred from high prices today that prices are more likely to be higher tomorrow. But it will typically be the case that this inference is tempered by other factors which are unrelated to information about tomorrow's prices. For example, there may be an unobserved shock to consumption demand for wheat today, which has shifted out the demand curve and pushed current prices up. Or buying by informed traders may be higher than expected because informed traders are less risk averse than expected (which will actually tend to *depress* tomorrow's price).

In the literature, one finds two sorts of models of rational expectations. In the more complex type of model, prices depend both on the information held by informed agents (through the impact of the information on the actions of the informed agents) and on other factors such as demand shocks, the level of risk aversion of informed agents, and so on. Prices in this sort of model are signals about privately held information, but they are necessarily noisy signals. The term *noisy rational expectations equilibrium* is used for models in which equilibrium prices depend on factors other than private information about which the uninformed are uncertain.[1] In the literature, the typical model simply assumes that there is some unobserved random component in aggregate supply. When the model is one of a securities market, the story most often told is that this represents net supply from liquidity motivated traders; these liquidity motivated traders are called "noise traders."

In contrast, in the early models of rational expectations agents know all the factors that determine price except for privately held information. Hence an uninformed agent can make a relatively direct inference from observed prices to privately held information. In this sort of model, current prices *may* provide all of the relevant information.[2] For lack of a better term, I will refer to this sort of model as one of a rational expectations equilibrium *without noise.*

---

[1] This is not quite precise as stated. The existence of some other factor besides private information that drives price(s) is neither necessary nor sufficient for prices to be noisy indicators of information. If, for example, all sources of uncertainty are drawn from discrete distributions, then it will "typically" be the case that every possible equilibrium price is associated with a single array of random factors, and equilibrium price reveals private information perfectly; see Radner (1979). But in the models investigated in the literature, sources of randomness in prices other than private information typically render prices a noisy signal of the private information that is relevant to the uninformed. The term *noisy rational expectations equilibrium* should properly be reserved for equilibrium prices that are noisy signals of the relevant private information.

Two fundamental problems with the concept of a rational expectations equilibrium should be noted. First, how is it that agents in the economy are able to disentangle the precise relationship between prices and private information? From where do their rational expectations come? To believe in the powers of rough introspection is one thing, but to believe that those powers are so substantial as to give perfect contingent insight is quite another, especially when it comes to endogenous variables such as equilibrium prices. A more believable story than the story of pure and perfect introspection, but one not easy to flesh out, is that agents in an environment which is "stationary" will come, by experience, to understand the relationship between prices and contingencies. Some interesting work has been done on this question (see, for example, the introductory paper by Blume, Bray and Easley (1982) and the references given there), but, after consuming this literature, still one must engage in a leap of faith in accepting rational expectations as an exact model. Perhaps the answer to this problem lies in a result that would establish that models based on expectations that are precisely rational are reasonable approximations to situations where expectations are "pretty good" if not perfectly rational. A general formalization of this sort of robustness result is unknown to me, although the mechanics of Anderson and Sonnenschein (1982), in which the information contained in prices is only imperfectly perceived by agents, seems a good place to begin for someone interested in putting together such a story.

The second fundamental problem with the story of rational expectations equilibrium is that, in most of these models, current equilibrium prices convey information that affects the demand and supply decisions of some of the agents for the good whose price is in equilibrium. But how can agents condition their demand on a price which is in part determined by that demand? In other words, is there any mechanism or market institution that could implement a rational expectations equilibrium? There are ways that it might be done. Imagine that agents submit entire demand correspondences, which implicitly include calculations like "if such and such is the price, then I infer this, and my demand (accordingly) is this. But at a different price, I infer that, and my demand is that."[3] Although some (weak) informal attempts in the literature have attempted to interpret limit orders as mimicking the submission of a demand correspondence, this mechanism doesn't resemble most of the institutions we observe. Real questions about the implementation of rational expectations equilibria remain to be answered.

---

[2] It is also possible that prices do not allow agents to infer all that is relevant in privately held information. For example, if a sufficient statistic for all private information is more than one dimensional and only a single price is being observed, then it would be impossible for the price to reveal the sufficient statistic.

[3] Whether this mechanism will, in fact, work, depends on the extent to which equilibrium price is a sufficient statistic for agents' private information. See Beja (1976).

Beginning with his thesis in 1975, Grossman was one of the earliest practitioners of rational expectations equilibrium. He contributed to the development of the literature on this topic on many fronts, not the least of which was putting it on the front burner of consideration by economists. More specifically, one can point to at least three issues on which he made significant contributions:

The first concerns the questions: How much information will be produced (privately) in a market economy, and how much information can be conveyed by equilibrium prices? Hirshleifer (1975) argues that there may be a socially wasteful overproduction of information. He asks us to consider a case in which information has no social value—it neither improves production nor allocation decisions—but where information may be valuable privately, because it enables those who have it to trade advantageously with those who do not. In [3] and elsewhere, Grossman (with Stiglitz) argues that Hirshleifer has failed to take into account the fact that, with rational expectations, prices may reveal private information to such an extent that one cannot gain privately from privately held information.

To take an extreme case, imagine an economy with identical agents holding identical initial endowments and having identical strictly convex preferences, so that no trade is possible which will benefit one person without hurting another. In Hirshleifer's model, a private individual might still seek to acquire private information —say information about which states are the more likely—and then buy up claims to consumption in those states. Grossman argues that this will bid up the price for claims to consumption in those states, revealing the privately held information. In fact, if all agents have rational expectations, then no trade can ever take place. To see this, note that in a rational expectations equilibrium, all parties must finish (in terms of their overall or ex ante utility) no worse off than they start. Since we began at a Pareto efficient allocation, the only trades that are possible in this setting of rational expectations are those that leave all traders at their initial level of utility, which means no trade at all given our assumptions. Put another way, to undertake a trade in this special environment, a trader must either be a fool or must suspect that the other side to the trade is a fool. In a rational expectations equilibrium, neither is possible, and there can be no trade.[4] But if no one will ever trade in a rational expectations equilibrium, then privately held information can be of no value. And then private information, if it has any cost at all, will not be produced.

This is clearly an extreme example. Suppose that, even in the absence of private information, some trade would take place in an amount that uninformed agents cannot predict perfectly beforehand. This could trace from any of the sources we discussed above such as unobserved shocks to demand or supply, or uncertainty about the level of risk aversion of other traders. Then the unknown trading levels provide the informed agents with a "screen of noise" behind which they may disguise their informationally motivated trade. The uninformed agent, faced with a given level of demand from the rest of the economy, can't be certain that this demand results solely from the actions of informed traders who have superior information. When informed traders can hide behind this sort of noise, they can benefit privately from private information, even in a rational expectations equilibrium. The private value of that information will depend on how much of it leaks out in equilibrium prices, or, put another way, on how well the informed can hide behind the noise that is present. Going back to Hirshleifer's argument, it is therefore not at all clear that there will be social overproduction of information with rational expectations. Indeed, there can be cases where information has social value (because it is useful in directing production efforts, for example), but since there is very little noise behind which the informed can hide, less than the socially optimal amount of information will be produced.

---

[4] The general result behind this simple example, known as the "no-trade theorem" in the literature, can be found in varying degrees of generality in Stiglitz (1971), Kreps (1977), and Milgrom and Stokey (1982).

In [13], Grossman and Stiglitz go on to propound a corollary to this. If by an informationally efficient market we mean (following Fama, 1970) that equilibrium prices "fully reflect all available information," then the private value of information in an informationally efficient rational expectations equilibrium will be zero. An individual agent who might become informed can just as well observe the equilibrium price and obtain for free a sufficient statistic of what would be costly to obtain privately.[5] If information has zero private value in equilibrium, but it does have a cost, then no private information is produced. If no private information is produced, then prices will not reflect the potentially available information. But then if agents are competitive (no single agent can affect equilibrium prices), there may well be private gains to any single agent from the information. Assuming that the cost of information is small enough, there can be no overall equilibrium with endogenous information acquisition, at least when prices would be informationally efficient. Hence the title of [13]: "On the Impossibility of Informationally Efficient Markets."

What saves equilibrium are circumstances that make equilibrium prices less than perfectly informationally efficient because of the sort of noise discussed above. In [13], Grossman and Stiglitz go on to construct a model of a noisy rational expectations equilibrium in which the decision to become informed is endogenous. This model is then used to refine our intuition on a number of important comparative statics propositions concerning, for example, the informativeness of prices, the number of traders who choose to become informed, and the volume of trade.

Related to these papers is Grossman's second specific contribution to the theory of rational expectations, contained in [7] and [22], concerning the relationship between the amount of information conveyed by equilibrium prices and the number of markets that are open. Recall the example of the spot market in wheat. Imagine that the equilibrium spot price of wheat provides an imperfect indication of the information received by informed speculators about tomorrow's spot prices because of noise in (say) current consumption demand for wheat. Now suppose that a futures market is opened today, in which bushels of wheat are traded, deliverable tomorrow. An agent who has private information about tomorrow's spot price of wheat will not only use that information to decide how much wheat to store, but also to speculate in the futures market. Both today's spot price and today's futures price will rise. But whereas today's rise in spot price might be confounded with shocks of today's consumption demand, the rise in today's future market price will not be. Even if there is other noise confounding the information contained in the futures market price, one expects in general that the information contained in the two prices will provide more insight into private information than will the spot market price alone. A model that works out these general points and refines them is developed in [7].

---

[5] It should be noted here that this argument relies heavily on the unspecified process by which equilibrium prices and net trades are set. There is an implicit assumption that a single agent can wait until after the equilibrium price is set before choosing how much to demand or, if not that, that the market mechanism allows the agent to condition his/her trade on the equilibrium price. In Dubey, Geanakoplos and Shubik (1983), for example, a trading mechanism is specified in which uninformed traders cannot do this. In this case one finds an equilibrium where the "prices" do reveal private information perfectly, and yet private information has positive private value. Of course, interest attaches to the concept of an informationally efficient market because one supposes that uninformed agents can take actions contingent on equilibrium prices, which is consistent with the argument in [13]. Still, it is especially desirable in the context of this issue to understand better the micro-structure of the market.

Also in [7], Grossman formulates a conjecture concerning why it is that we don't see a complete set of markets (in the spirit of Arrow and Debreu). The argument is that the more contingencies for which there are markets, the more prices there are and, therefore, the more private information will be revealed by equilibrium prices. This will drive down the private value of information, and it will lessen the amount of private information that is produced. "Hence," Grossman writes, " 'speculative capital' would not be attracted to many markets and they would thus be too 'thin' to exist."

The implicit presumption in this argument is that very complete markets will be very revealing of private information. In [22], Grossman puts this presumption on sound theoretical footing. He asks there: What information will be transmitted by equilibrium prices when there is a complete set of (payoff relevant) contingent claims markets? The answer he obtains (under some mild technical restrictions) is that, for all practical purposes, all useful information can be conveyed. More precisely, if one looks for the equilibrium prices and allocation that would result if all private information was made public, then, with complete markets, that equilibrium is a rational expectations equilibrium for the economy with privately held information.

Grossman's third specific contribution concerns rational expectations equilibrium prices that aggregate diverse publicly held information. In many of the early models of rational expectations equilibrium, the model imagines two types of agents—those without information and those with. Moreover, those with information all have the same information. Hence the issue becomes the extent to which the equilibrium price(s) conveys the single piece of private information that is available. One imagines, however, that different privately informed individuals have access to different pieces of information. The issue then concerns not simply the (noisy) transmission of a single piece of information, but instead the way in which equilibrium prices aggregate the wealth of available information. In [3], [4] and [10], Grossman studies examples of this type of information aggregation. In [3] and [4], parameterized models are worked out which give the conclusion that there can be cases in which prices provide a very good signal (a sufficient statistic) for all the information that is in the hands of private individuals. In [10], the analysis of [3] is generalized to a large extent—roughly put, Grossman provides here the foundations for extending the capital asset pricing model of Lintner (1965) and Sharpe (1964) from cases where all information is public to cases where some of the information may be in private hands. These early models of rational expectations equilibrium prices that aggregate information do not encompass the sort of noise that allows us to avoid the Grossman-Stiglitz "informativeness paradox," but they inspired later work which continues to this day, advancing our understanding of markets in which some agents hold private information and all agents try to learn from prices what others know.

Issues Concerning the Firm

One theme that has played consistently in Grossman's research, albeit in a number of different ways, concerns the structure of the firm and decision making by the firm. In this section I will survey four variations that he has played around this general theme (for the most part in collaboration with Oliver Hart): (1) the objective function of the firm when markets are incomplete; (2) takeover bids and their effect on the managers of firms; (3) incentive theory and contracts; and (4) a transaction cost analysis of patterns of ownership.

Incomplete Markets and the Objectives of the Firm

In classic general equilibrium theory, firms maximize profits. As long as firms are competitive (that is, their choice of production plan doesn't affect prices), this leads to a choice of production plan that all shareholders applaud. This is self-evident: A shareholder's utility can only be enhanced if his opportunity set is made unambiguously larger, and a competitive firm which maximizes profits maximizes the shareholder's wealth to the extent that it can without being able to change the relative prices facing the shareholder.

Two complications arise when time and uncertainty are added to the story, and when we envision a sequence of markets as in Arrow (1953) and Radner (1972), rather than a once-and-for-all set of contingent claims markets, as in Debreu (1959). First, shares of stock in the firm become traded securities; if the firm's ownership changes from moment to moment, it is conceivable that the objectives of the owners at one date do not agree with the production plans that were favored by owners at an earlier date. Second, markets may be incomplete, so that a firm may change the range of consumption possibilities open to consumers by changing its production plan and (therefore) the pattern of returns that its shares will generate.

A simple example will illustrate these points. Imagine that there are two dates in this economy, zero and one, and at date one there are two possible states of nature. There is a single consumption good, say, wheat. All firms are identical and rather trivial: A firm can store a unit of wheat at date zero and take that unit of wheat out of stock at date one. The firm can also destroy some of the wheat, if it chooses to do so. Only firms are able to store wheat. Markets in wheat and shares of the firms are open at date zero. At date one, since the shares of the firm have no continuing value and there is a single commodity, there is no need for any markets. Firms pay dividends in the form of the wheat they have stored (and not destroyed), and consumers eat however much wheat they possess. It is clear that a firm will maximize profits only if it does not destroy any of the wheat it has chosen to store. But we shall see that, in some circumstances, shareholders would *want* to see a bit of wheat destroyed so that, by trading in the shares of the companies, they can create insurance for themselves.

Imagine that we have two types of consumers, classified according to their endowments. One type of consumer is endowed with, say, one hundred bushels of wheat at date zero, one hundred bushels of wheat at date one in the first state, and zero bushels of wheat at date one in the second state. The second type of consumer has the same date zero endowment, but her date one endowment is the mirror image of the first consumer's: zero bushels in the first state, and one hundred bushels in the second. Consumers all have preferences which lead them to wish to level their consumption through time and across states as much as possible. There are equal numbers of consumers of each type. In addition to their commodity endowments, every consumer holds (ex ante) a pro rata share of every firm in the economy. Firms, in the standard traditions of general equilibrium, hold no endowments ex ante.

If markets were complete, consumers would trade away from their endowments to reach the position where they consume seventy-five bushels at each date, in each state. Note well that this requires the simple production function of the firms—twenty-five bushels of wheat per consumer must be stored.

We suppose, however, that there aren't a complete set of markets. The only way that a consumer can save from date zero to date one is via his or her shares in the firm. Shares at date zero will have positive value for firms that store wheat, since these shares will pay off in wheat-dividends at date one.[6] The problem is that there is no way for either type of consumer to save more or less wheat for one state versus the other. As long as firms maximize profits, shares pay the same dividend in the two states. Since holding shares in firms is the only possible way to save in the economy, the riskiness in consumer endowments will result in risky consumption at date one for all consumers.

Now imagine that shareholders in one of the companies, noting all this, propose that the company engage in the following production plan. Store exactly one bushel of wheat. In the first state, distribute the one bushel pro rata to the (then) shareholders. In the second state, destroy the one bushel. A share in this company, then, is an Arrow-Debreu contingent claim to the consumption of wheat in the first state. Moreover, if there is another firm in the economy that is storing, say, N bushels of wheat and destroying none of it, then by buying a $1/N$ share of this second firm and selling short a share of the first firm, one has a package of shares that pays a zero dividend in the first state and a positive dividend in the second. That is, this package is an Arrow-Debreu contingent claim to consumption in the second state. Hence, once our first firm adopts this wasteful production plan, consumers, by trading in the shares of firms, have available a complete set of contingent claims markets, and they can insure perfectly against their risky endowments. Because some resources are being destroyed, we cannot get quite to the efficient outcome obtained if there were originally a complete set of contingent claims markets. But only one bushel of wheat is destroyed and only in one of the two states, so we get very close.[7]

---

[6] The initial capitalization of the firms, used to purchase the wheat they store, is presumed to come from initial shareholders.

Note well: As long as the need for insurance by shareholders is sufficiently great, (initial) shareholders are unanimous in agreeing that they are better off if the first firm commits to this wasteful production plans than if it doesn't; without some wastage by some firm, there is no way to insure against random endowments. But if we imagine some consumers who are without any need to insure across the two states but who hold shares in this firm, we will destroy this unanimity; since these shareholders have no insurance motive, they see no reason to waste any wheat at all. Also, when date one rolls around, if we are in the second state of nature, the (then) shareholders of this firm will see no point in carrying out the destruction of the bushel of wheat (which only serves to lower their dividend). So unless the initial shareholders in this firm can commit the firm to destroy the wheat in the second state, no one (with rational expectations) will believe that this plan will really be carried out.

This simple example is not very realistic, but it makes the point very starkly. When shares in the firm are traded, markets are incomplete, and unbounded levels of short selling is permitted, the firm can greatly change (in this case increase) the market opportunities of consumers by undertaking a small change in its production plan. Even if this small change decreases the wealth of shareholders in the firm, the shareholders may be more than compensated by the change in available markets. Of course, in an example this stark one is left to wonder why markets aren't complete to begin with. Market incompleteness must be traced back to some form of transaction costs; since the model above (and models of the sort that populate this literature) lacks any such costs, it cannot provide an answer. It would be better to have models in which market incompleteness arises endogenously. Still, insofar as markets are incomplete and firms may change the feasible consumption set when they change production plans, the effect noted here remains of interest.

In the early 1970s a number of authors attacked the general problem of objectives for the firm when markets are incomplete and shares trade in a stock market. (See especially the 1974 symposium in the *Bell Journal of Economics*.) These authors concluded that problems with the classic criterion of profit maximization arise because of a lack of "spanning," defined as follows.

Fixing a set of production plans by firms, agents are able to use the markets that are open or will be open, including all the financial securities that are available (including shares in the firms) to make various net trades. Assuming that there are no transaction costs and no short sales constraints (which is standard in this literature), the set of feasible net trades forms a subspace of the entire commodity space. We refer to this subspace as the *span* of available markets. Note well that as firms change their production plans, they may change the span of available markets because their shares are financial securities. At an equilibrium, the "spanning property" is said to hold if any contemplated deviation in the production plans of a single firm will not change the span of available markets. For example, if some exchange in the economy offers a complete menu of Arrow-Debreu contingent claims, then markets will be complete no matter what production plans any firm adopts, and the spanning property holds.

---

[7] By scaling down the operations of the first firm, we get asymptotically close to the efficient equilibrium outcome.

The literature referred to above concludes that if the spanning property holds, (final) shareholders unanimously agree that firms should "maximize" profits. The word maximize is in quotes because the firms are meant to disregard any effect they have on prices and take the prices as given. From the point of view of classic theory, this is surprising. In, say, a two-period, one state of the world economy, where there is a complete array of spot markets at each date and some way to save from the first period to the second, the spanning condition will hold. Yet we are used to the conclusion that shareholders (unambiguously) want firms to maximize profits only when the firms are competitive; this literature seems to say that the assumption of competitive firms is unnecessary.

The literature comes to this surprising conclusion because it is assumed (in this literature) that there are two periods, with (perhaps) many states of nature in the second period, but only a single consumption good. Since there is nothing to trade for in the second period, initial shareholders in a firm will "eat" any output that the firm produces in proportion to their shareholdings. Accordingly, if a firm has and uses market power, it does so to the detriment of its shareholders. In [5] and [14], Grossman and Stiglitz make this point, and they show how, in models with (for example) three periods, it is incorrect that shareholders unanimously favor profit maximization if the spanning condition holds. The same point can be made if there is more than one consumption good in a two-period model. They go on to show that the twin assumptions of spanning and competitive firms *will* guarantee that shareholders unanimously concur with profit maximizing production plans, but that spanning will be a problematic assumption when markets are incomplete and firms are able to change their debt-to-equity ratio.

What if spanning fails, and/or firms are not competitive, so that there is no agreement (necessarily) among shareholders about what firms should do? In [9], Grossman and Hart propose an objective that the firm might have, based on a story of shareholders meetings that runs as follows: If a production plan is proposed that lies outside the span of what is currently marketed, every shareholder must first decide how the market will value that production plan. Grossman and Hart assume that shareholders have "competitive conjectures;" that is, each shareholder assumes that the market will value the plan (on the margin) exactly as the shareholder would do if the shareholder were forced to consume the plan.[8] In general, these conjectures will vary from shareholder to shareholder, and we imagine that, at the shareholders meeting where a particular plan will be adopted, monetary bribes flow between shareholders to arrive at a consensus. That is, if (on the margin) my personal opinion is that plan X is going to be very profitable, I might be willing to buy the votes of other shareholders who are not so optimistic. But I am also willing to sell my vote for plan Y, if the bribe I get more than compensates me relative to what I suppose will be the fall in the value of my shares if Y is adopted instead of X. In an equilibrium of this bribery process, the production plan that is adopted is one that maximizes a weighted average of marginal valuations of shareholders, with weights related to shareholdings. Grossman and Hart go on to establish an efficiency characterization of the equilibria of economies where firms have this "objective function."

---

[8] Of course, when the plan falls within the span of things already in the market, the shareholder's private marginal valuation is aligned with market prices, at least if the shareholder is at an interior solution in his consumption planning problem. So in this case we get back the "unanimity" result, because shareholders think the firm is competitive, even if it is not.

Takeover Bids

In the literature just described, one doesn't worry that managers of the firm will fail to listen to shareholders, at least if shareholders speak with a clear enough voice. But let us make sufficient assumptions so that shareholders do have clear and unanimous objectives; assume that the spanning condition holds and the firm is competitive. There is still the question of whether management will put into place the production plan that maximizes profits.

The literature is filled with reasons why we might doubt that management will maximize profits, many of which come down to: Management might have its own objectives to pursue. Of particular interest for what is described next is the case where current management may not be the most efficient, but they would rather stay as managers than be replaced by someone more efficient. One of the many mechanisms that is meant to provide discipline for managers, so that they do act to maximize profits, is the corporate takeover. If the managers of a firm are inefficient (either because they willfully choose an inefficient production plan or because they simply aren't as talented as some other managers), then a more efficient outsider is meant to be able to take over the firm, benefiting by the raised profits of the firm (Manne 1965; Marris 1964).

In [12], [16], [17] and [22] (and in ongoing work today), Grossman and Hart have explored this simple story concerning takeovers, and they find it wanting. The fundamental observation that they make is that there is a free-rider problem in takeover bids. Imagine that a firm is inefficiently run, and an outsider announces a tender offer for its shares. An individual shareholder realizes that there are two possibilities if the takeover succeeds. Either the shares will be worth more after the takeover than the tender offer price, or they will be worth less. In the first case, there is no reason to tender shares—better to free-ride on the efforts of the party who is tendering (the raider). The second case is hard to believe: Why would the raider tender for shares at a price greater than what they will be worth if the takeover succeeds? Only if one believes that the raider is trying to lose money (or, perhaps, is expecting the raid to fail, but is hoping for greenmail) can one believe that the shares after the takeover will be worth less than the tender price. So the only sensible thing to do in a rational expectations equilibrium is to free-ride on the tender offer and refuse to tender. This implies that the takeover cannot succeed (unless there are many shareholders who are not taken with this logic). So if takeover bids are costly, takeovers will never be attempted, and the takeover mechanism is not an effective discipline on management.[9]

Of course, there are implicit assumptions in the argument above which are not necessarily valid. First, if the tendering party (hereafter, the raider) already holds an interest in the target company, the raider will benefit from the increased worth of those shares that he already holds. To increase the worth of those shares, the raider may be willing even to pay a small premium above the (marginal) value of shares to him, in order to induce other shareholders to tender. This explanation for takeovers is given in passing in [17] (see fn. 5) and is rejected there; but this story has been picked up more recently by Shleifer and Vishny (1986). Grossman and Hart focus on a second implicit assumption: The argument assumes that the value of a share to a shareholder who doesn't tender is the same as the value of a share to the raider. But suppose that the raider can dilute the value of the shares held by minority shareholders. For example, the raider may be able to sell assets of the firm to another firm owned by the raider at an advantageous price, or he may be able to consume some of the firm's resources directly as perks, or he may be able to direct the actions of the firm in a way that benefits some other enterprise that he owns. If there is a gap in these values, then it is possible that the raider will tender at a price below the value to him but above the value to minority shareholders (if the raid succeeds); so that minority shareholders who believe that the raid will succeed will be willing to tender.

Grossman and Hart conclude then that the ability of a raider to dilute the value of shares of minority shareholders is crucial to the discipline of the takeover mechanism.[10] If no dilution is possible, then no takeovers will take place. In writing the charter of a new corporation, shareholders will wish to put into place some possibility of dilution, if they want this discipline to act on their managers. In the model of Grossman and Hart initial shareholders choose a level of permissible dilution to write into the corporate charter that balances the following three forces: (a) a greater possible dilution increases the chances of a takeover, which (in the equilibrium described by Grossman and Hart) raises the value of shares above the pretakeover level; (b) in addition, if there is a greater threat of takeover, and if incumbent management doesn't wish to be replaced, the greater threat of takeover will cause management to be more efficient (insofar as it is able) in choosing production plans ex ante;[11] (c) but greater possible dilution reduces the tender price in the event of a tender offer, at least when the raider faces no competition in the takeover attempt. In [12] and [17], these points are all made and the trade-off just described is analyzed. Grossman and Hart also look at questions concerning the socially optimal level of dilution. Since part (c) of the tradeoff only concerns the redistribution of wealth between original shareholders and the raider, they find that the socially optimal level of dilution is generally greater than the level that would be selected by original shareholders who are composing a corporate charter. And in [16], Grossman and Hart study this question with particular concern paid to the provisions of the Williams Amendment to the Securities Exchange Act of 1934.

---

[9] If there is some uncertainty about the value of the shares, even if the raider possesses superior information about the value, a similar argument can be made, as long as we assume that shareholders and the raider would agree on the value of shares of stock, given the same information. All these arguments depend on a maintained hypothesis that each shareholder assumes that his own decision of whether to tender will not affect the success of the takeover attempt.

[10] If we rely on the other explanation for takeovers, that raiders can acquire a small holding in the company before coming forward with a tender offer, then it is crucial that potential raiders not be forced to reveal themselves too quickly.

[11] Although outside the model of Grossman and Hart, one can imagine that some activities by incumbent management that are meant to forestall the possibility of a takeover, such as poison-pill provisions, will work the other way.

In [21], Grossman and Hart add two important elaborations to the basic model described above. First, the earlier models consider the case where, at the time the tender offer is made, all parties are aware of the value of shares to minority shareholders after a successful raid. This assumption is replaced by the more reasonable one that the raider has superior information about this. Even with this, they find that the basic conclusions from their previous work hold up, as long as shareholders have rational expectations. Second, they consider an additional rationale for attempting a takeover, other than to improve the performance of the firm: The raider may have superior information about the prospects of the firm under incumbent management. Raids based on this motivation are purely redistributional; Grossman and Hart call these *acquisitional raids*, as opposed to *allocational raids*, which are based on increases in efficiency to be effected by the raider. Suppose an acquisitional raid is attempted, and it fails. The very fact that the raid was attempted would indicate that the firm is worth more under incumbent management than was previously supposed. Hence we would expect to see the value of firms increase (on average) after failed raids.[12] Also, if a raid is identified as being acquisitional, it cannot succeed: Since the firm has been identified as being worth more under incumbent management, and the raider will dilute the value of minority shareholdings to compensate himself for the cost of the raid, shareholders are better off with incumbent management. This is a simple variation on the no-trade theorem. In any raid that is clearly acquisitional, there are no social gains from trade. The only way that a raider can gain is if current shareholders lose. Only when some raids are allocational, and acquisitional raiders can "hide behind" the possibility that their raid is allocational, can acquisitional raids succeed at all. Finally, Grossman and Hart speak to the policy issue: Acquisitional raids are socially undesirable since they are purely redistributional, except for the resources they consume in research and in making the tender offer. Perhaps, therefore, when takeovers can be either acquisitional or allocational, there is cause for the government to restrict raids in general, say, by regulating the amount of dilution that is allowed. Within the context of their model, Grossman and Hart demonstrate that this is bad policy. The socially optimal level of dilution is always greater than the level that would be selected by initial shareholders.

Grossman and Hart continue to pursue this topic. In an unpublished paper, "One Share/One Vote and the Market for Corporate Control," they study how effective the takeover mechanism will be for corporations that have multiple classes of stock (for example, one class of dividend accruing stock, and a second class of voting shares).

---

[12] For empirical evidence, see Dodd and Ruback (1977).

Incentives in Firms and Incentive Theory

Takeover bids are but one of the many forms of incentive device for managers. There are certainly many others. Grossman (with others) has studied some of these in specific contexts. In [15], for example, Grossman and Hart explore financial structure and bankruptcy as an incentive device. In this paper, they consider the manager of a firm whose sole objective is to maximize the utility of the perquisites (a pleasant office, a fancy corporate plane, corporate retreats in Tahiti, and so on) that the manager consumes. The manager takes the initial financial assets of the firm and chooses how much to invest in a (potentially) profitable investment and how much to spend on perquisites. The manager enjoys the perquisites only if the firm does not go bankrupt. If the firm was financed entirely by equity financing, there would be no chance of bankruptcy, and the manager would spend all the firm's capital on perks. Accordingly, potential investors in an all equity firm would rationally anticipate that none of the initial capital of the firm would go into the profitable investment and would refuse to invest in it; the cost of capital for this firm would be infinite. However a firm which is financed in part by debt would be unable to repay that debt if all its capital went for perks. So if (as Grossman and Hart assume) bankruptcy causes the manager to lose the value of his perks, the addition of some debt financing effectively creates for the manager some incentives to invest profitably some of the firm's capital. At every level of debt-to-equity and total working capital, investors can determine how the manager will allocate that capital in his own best interests; how much into profitable investments (to lower the chances of bankruptcy and loss of his perks), and how much for perks. Investors can then determine the return they can expect on their investment given the decisions of the manager. And this determines how much they are willing to loan. For every level of the debt-equity ratio, we find equilibrium amounts of capital that can be raised. (For example, at a zero debt-equity ratio, no equity is raised.) And the manager sets the debt-equity ratio to maximize his own utility. Note well that the total amount the manager will be able (and willing) to take out in perks rises (at first) with increasing debt-to-equity, since only with some debt will anyone invest at all in the firm.

Grossman, with Hart in [20] and [28] and with Hart and Maskin in [32], also employs the general methods of incentive theory (in cases akin to market signaling) to study questions of the use of "inefficient" unemployment by firms. The basic idea is that, while it may be efficient for firms to raise or lower the wages of workers as the firm's profit opportunities improve or decline, if only the firm sees the level of these profit opportunities, the firm will always want to claim that times are bad for it so that it can lower wages. Raising or lowering the level of employment to inefficient levels is a credible (incentive compatible) signal the firm can send concerning its profit opportunities, which provides some insurance to workers. Papers [20] and [28] study the problem from the perspective of the individual firm; in [32] this is integrated into a general equilibrium treatment.

Owing to lack of space, I will not review these papers further here. Instead, I will discuss the methodological contribution to incentive theory by Grossman and Hart [26]. The classic incentive problem has two actors, a principal and an agent. The principal seeks to hire the agent to perform a certain task, but the principal cannot monitor the amount of effort that the agent exerts. The agent is effort averse, so that he would exert minimal effort if his compensation was insensitive to his effort. The principal is able to make the agent's compensation depend on a measure of the agent's level of effort, namely the level of output obtained. But output is a imperfect indicator of the level of effort since it depends not only on effort exerted, but also on random factors that are outside the control of the agent. If the agent's compensation is tied to the level of output he produces, the agent is subjected to a risky level of compensation. If a principal is risk neutral and the agent is risk averse, then shifting any risk onto the agent is inefficient. This is an unavoidable inefficiency, however, if the agent is to supply more than a minimal level of effort. The questions then are: Given the incentive problem, what is the optimal compensation contract to offer the agent? What properties will this contract have? How much effort will the agent exert at this "second best" contract? How will the contract, effort levels, and the principal's net level of profit change with changes in various parameters of the problem?

Many authors wrote on this problem prior to Grossman and Hart. (See, for example, the bibliography of [26].) The standard method of analysis had become the so-called first order approach. Essentially, the problem of finding the optimal incentive scheme is set up as a constrained optimization problem, where one is seeking to find an optimal compensation scheme, subject to the constraint that the agent is willing to participate in the scheme at all, and subject to a constraint that is meant to capture the idea that the agent's choice of effort is best for the agent, given the compensation scheme. The rubric "first order approach" refers to the formulation of the second constraint: It is assumed that the agent's choice of action is whatever solves the first order equation for the agent's optimization problem. Mirrlees (1975), however, showed that the first order approach does not generally guarantee a solution, nor does the optimal solution necessarily satisfy the first order conditions.

In [26], Grossman and Hart provide an illuminating way for attacking the general problem. To use this technique, it is necessary that the agent's disutility of effort is separable from his utility from compensation; in this discussion I will assume that the agent's overall utility takes the very simple (but standard) form of u(compensation) $-$ v(effort level), for some functions u and v. To keep matters simple, we will follow the exposition of [26] and assume as well that the principal is risk neutral. (This second assumption is not necessary, as Grossman and Hart point out.) Then for every effort level that the agent might take, it is easy to compute the benefits to the principal if the agent chooses that action *gross* of any payments made to the agent. This is just the expected profit that the agent's activities will generate, as a function of the agent's effort. Let us denote by e the level of the agent's effort, and let $B(e)$ denote this gross benefit function. It is relatively easy to derive properties of the function $B(e)$ from primitive assumptions on the problem.

Now fix some level of effort e, and ask: What is the least cost at which the agent can be induced to exert effort level e? If we think in terms of a finite number of possible output and effort levels, this can be cast as a convex programming problem with linear constraints. Let $\{e_1, \ldots, e_n\}$ be the possible effort levels, and $\{x_1, \ldots, x_m\}$ be the possible levels of output. The effort level e that we are trying to induce is some one of the e_j; let us denote it by e_J. For decision variables, take the levels of gross utility of compensation obtained by the agent at each outcome x_i, denoted (say) u_i. Now each level of effort induces a corresponding probability distribution on outcomes, so the condition that effort level e_J is chosen is easily written (in terms of the decision variables u_i) as a set of linear inequalities. For example, if p_{ij} denotes the probability of output i if effort level j is taken, then the constraint that e_J is preferred to, say, level e_1 is simply

$$\sum_{i=1}^{m} p_{iJ}u_i - v(e_J) \geq \sum_{i=1}^{m} p_{i1}u_i - v(e_1).$$

(The constraint that the agent's expected level of utility is sufficient to ensure his participation in the scheme has the same left hand side as above and a constant on the right hand side.) Moreover, we can invert the agent's utility function to determine what level of income he must be given to obtain gross utility u_i at outcome x_i. Let us write this level of income $w(u_i)$; it is the solution of $u(w) = u_i$. Then our problem is to

$$\min_{u_1, \ldots, u_N} \sum_{i=1}^{N} p_{iJ}w(u_i),$$

subject to the linear constraints outlined above. If the function $u(\cdot)$ is concave, then $w(\cdot)$ is convex, and we have a relatively simple convex programming problem. Even if e is a continuous variable, so that we have a continuum of linear constraints, this problem is relatively simple.

For effort level e, let the value of the objective function (the minimum level of cost) be written $C(e)$. Then the final stage of the problem is to choose e to maximize $B(e) - C(e)$.

So what? In the first place, the decomposition of the principal-agent problem (into the computation of $C(e)$ for every e, and then the determination of the optimal e) is extremely useful pedagogically. At least, I have found it so in the classroom. It sets the stage for a useful existence result (Proposition 1 in [26]). It can be used rather naturally to generate some useful conclusions about the shape and behavior of the optimal compensation schedule, because it separates the determination of that scheme from the determination of the optimal level of effort to induce. That is, whatever level of effort e^* maximizes $B(e) - C(e)$, the compensation schedule that goes with it is the solution to the simple convex programming problem for $e = e^*$. If one can obtain characterizations of the solution to that problem for every level of effort e, then those characterizations hold good for the compensation schedule that induces level e^*. Grossman and Hart go on in [26] to do just this exercise, giving, for example, conditions under which the optimal compensation scheme is increasing in the level of output, under which the optimal compensation scheme is progressive (or regressive), and so on. In the case of two possible output levels, it allows for a very complete analysis of the problem (given in [26]). And it allows Grossman and Hart to extend Holmstrom's (1979) result on the use of "informative signals."

On Transaction Cost Economics

In [34], Grossman and Hart take yet another tack on issues concerning the organization of firms. In this paper, their approach comes from the field of transaction cost economics, as originated by Coase (1937) and developed especially by Williamson (for example, 1985). The issue is one of contractual incompleteness and governance provisions that are part of most contractual forms. Imagine two firms (an upstream and downstream firm, say, or two firms laterally related in some market), each of which must make some initial investment today. Tomorrow, further decisions must be made concerning the use of those initial investments, and then benefits accrue to the parties, depending on the pairs of investments chosen. For some unspecified reason, it is impossible to write an enforceable contract today which specifies levels of initial investment. It is also impossible to write contracts which specify what decisions will be made in the future, or to write a contract which depends on the future levels of the benefits received. While Grossman and Hart provide no single specific reason for this inability, they do cite the usual (fairly convincing) excuses: for example, these things are all too complex to be specified in a contract; or they are too hard or too expensive to verify in court.

Two types of contracts, however, *can* be written. First, the two parties can write a contract today concerning the ownership rights to the assets tomorrow. Essentially, one party can buy from the second the assets that the second will put into place. The second party still has free reign over what assets to put into place; all this party gives up is the right (tomorrow) to control those assets. This contract, of course, can be accompanied by a lump sum payment. Second, the two can write a binding contract tomorrow on the choice of action taken tomorrow. If they don't write such a contract, it is assumed that the choices of actions tomorrow are made noncooperatively by whoever controls the assets (as specified in any contract written today). It is assumed that the two will in fact write a contract tomorrow which is efficient ex post. The noncooperative solution is used only to decide how the surplus that cooperation will bring is to be divvied up. In their analysis, Grossman and Hart simply assume that the surplus above the noncooperative "threat point" is divided equally. The point of this is that given an assignment of ownership rights, investment levels today are made with at least half an eye towards strengthening one's bargaining position tomorrow. One anticipates an efficient outcome tomorrow, but one's share of that outcome depends on how well one would do if a noncooperative solution were to prevail. Ownership rights are (optimally) assigned so that the impact of this bargaining-driven inefficiency on the choice of investment level today is made as small as possible.

Grossman and Hart go on to analyze this situation under some mild technical assumptions. The exact results they obtain concern how ownership patterns affect initial levels of investment. These results are not at all surprising, and they will not be recounted here. What is more important than the specific results they obtain is the general nature of their analysis. It is apparent that contracts in complex situations are incomplete; that is, contracts routinely fail to specify what will happen in many contingencies. It is also apparent that negotiations and contract terms are often not about *what* will happen as contingencies arise, but what will be the *process* by which contingencies are met. That is, negotiations concern governance provisions. Finally, it is apparent that the greater framework of laws within which contracts are made impose on contracts certain implicit terms of governance. For example, ownership of physical assets gives the owner the right to command the use of those assets (more or less) as the owner wishes. While all these things are apparent, classical economic theory has not exactly bowed to them. In [34], Grossman and Hart provide us with one of the first mathematical models of these general ideas. In particular, they show how some patterns of ownership (which effectively delimit the boundaries of the firm) may be more or less efficient, as the patterns of ownership change the (implicit) governance mechanism and, therefore, the way in which transactions evolve as time passes.

And the Rest...

Of the forty published papers given in Grossman's c.v. of March 1987, I have touched on scarcely half (and some of those touches were exceedingly brief). Even so, to set the issues and give the briefest indications of Grossman's contributions, I have more than used up the allocation of space given to me by the editors. This is an indication of Grossman's phenomenal breadth as an economic theorist—in twelve years since his Ph.D. he has made very significant contributions to a great number of areas, writing in a number of styles, using a number of methodologies. I have selected from his work those topics that I know best, so I have selected primarily work in microeconomics, but there is work as well in macroeconomic theory, in [25], [29], [31], and [32]. His work on rational expectations and on takeovers pertains, of course, to financial markets. But his contributions to the theory of financial markets goes far beyond these works: he has contributed to the debate on stock price volatility as it relates to rational behavior in markets (in [20] and [27]) and to matters related to the administration of securities markets (in [33], [35] and [36]). Sequential bargaining is far from the mainstream of his interests, but [38] is a much cited paper there (and that study has led to an important contribution to the problem in noncooperative game theory of equilibrium refinements, [39], which is even further from his mainstream interests). In addition, [23] has become much cited, not only for its excellent economics, but also because it quietly initiates a boomlet industry on getting more out of signaling models by restricting out of equilibrium beliefs.

Grossman has tended to make his contributions at the level of contextual applications rather than methodological innovations. Methodological contributions are often made in passing, as in [23], [34] and [38], in the analysis of a specific contextual problem. This is not a general rule: [6], [26], [27] and [39] can all be

Table 2

The citations of Sandy Grossman

Paper number	*Year published*	*Total cites*	*Cites in set 1*	*Cites in set 2*	*Cites in set 3*
1	1975	21	2	5	1
2	1975	3	0	3	0
3	1976	84	11	14	16
4	1976	137	18	19	15
5	1977	15	2	1	3
6	1977	24	5	5	4
7	1977	76	7	23	6
8	1977	26	7	3	0
9	1979	26	3	8	3
10	1978	42	7	11	9
13	1980	131	15	22	10
14	1980	17	2	4	4
15	1982	8	2	0	1
16	1980	24	2	0	2
17	1980	55	4	0	14
18	1981	19	4	3	3
19	1981	65	16	9	1
20	1981	72	13	7	13
21	1981	8	1	0	1
22	1981	43	2	10	1
23	1981	25	4	3	2
24	1984	2	1	0	0
25	1982	19	7	1	1
26	1983	54	13	12	4
27	1982	18	0	0	7
28	1983	23	5	5	0
30	1983	2	1	0	0
31	1983	13	5	2	0
32	1983	9	4	4	0
33	1984	2	0	0	0
34	1986	6	3	0	0
37	1987	1	1	0	0
38	1986	6	3	2	0
39	1986	2	0	1	0
unpublished papers		14	5	3	2
Totals		1092	175	180	123

Citation counts for the publications and research papers of Sandy Grossman, taken from the *Social Sciences Citation Index*, 1975 to the first quarter of 1987, are given. Paper numbers correspond to Table 1. Only papers with at least one citation are given, and other unpublished papers which are cited are lumped together. Column 2 gives year of publication. Column 3 gives total number of citations. Column 4 gives total citations in the *American Economic Review*, the *Bell/Rand Journal of Economics*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics*. Column 5 gives total citations in *Econometrica*, the *Journal of Economic Theory*, and the *Review of Economic Studies*. Column 6 gives total citations in the *Journal of Finance* and the *Journal of Financial Economics*.

counted as solidly methodological. But Grossman does seem most at home in pushing at a particular economic problem in a specific context—he has an extraordinary talent for finding the economic nub and burnishing it for others to see.

His impact on the profession is as broad as his contributions. In Table 2, his forty published papers are again listed (by number—for titles see Table 1), together with citation counts through the first quarter of 1987, according to the Social Science Citation Index.[13] Column 2 gives year of publication (so subsequent data can be corrected for age, if the reader so desires), and column 3 gives total citations. Although it is completely unscientific, citations were grouped into three categories: Citations in the "mainstream" economic journals, the *American Economic Review*, the *Bell/Rand Journal of Economics*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics* (column 4); citations in the "theory" journals *Econometrica*, the *Journal of Economic Theory*, and the *Review of Economic Studies* (column 5); and citations in the two leading finance journals, the *Journal of Finance* and the *Journal of Financial Economics* (column 6). (Citations in other journals make up the balance.) Of course, many people write across these sets of journals; the sets of journals hardly speak to or for disparate groups. Still, one notes that Grossman's citations fall fairly among these three sets, which I think is indicative of his broad reach to economists of many stripes and even into the community of academic finance. He is surely a natural for the Clark medal, and the committee is to be congratulated for having chosen him.

■ *I am grateful to a large number of colleagues with whom I had many helpful conversations and even more enjoyable sessions reminiscing about Sandy. Specific comments on earlier drafts by Anat Admati, Mark Wolfson, and the editors were very helpful. The victim is in no way responsible for the preceding crime.*

---

[13] I am grateful to my research assistant Hideshi Itoh for digging out these data without complaint.

References

Anderson, Robert, and Hugo Sonnenschein, "On the Existence of Rational Expectations Equilibrium," *Journal of Economic Theory*, 1982 *26*, 261–78.

Arrow, Kenneth J., "Le Rôle des Valeurs Boursières pour la Repartition la Meilleure des Risques," *Econometrie, Colloques Internationaux du Centre National de la Recherche Scientifique*, 1953. Reprinted in English as "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Review of Economic Studies*, 1964, *31*, 91–96.

Beja, A., "The Limited Information Efficiency of Market Processes," Research Program in Finance Working Paper No. 43, Berkeley, University of California, 1976.

Bell Journal of Economics and Management Science, "*Symposium on the Efficiency of Competitive Capital Markets*," 1974.

Blume, Larry E., Margaret Bray, and David Easley, "Introduction to the Stability of Rational Expectations Equilibrium," *Journal of Economic Theory*, 1982, *26*, 313–7.

Coase, Ronald, "The Nature of the Firm," *Econometrica*, 1937, *4*, 385–405.

Debreu, G., *Theory of Value*. New York: John Wiley, 1959.

Dodd, P., and R. Ruback, "Tender Offers and Stockholder Returns: An Empirical Analysis," *Journal of Financial Economics*, 1977, *5*, 351–73.

Dubey, P., J. Geanakoplos, and M. Shubik, "Revelation of Information in Strategic Market Games: A Critique of Rational Expectations," Cowles Foundation Discussion Paper No. 634, Yale University, 1983.

Fama, Eugene, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, 1970, *25*, 383–417.

Hayek, F. H., "The Use of Knowledge in Society," *American Economic Review*, 1945, *35*, 519–30.

Hirshliefer, J., "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, 1975, *61*, 561–74.

Holmstrom, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, 1979, *10*, 74–91.

Kreps, David M., "A Note on Fulfilled Expectations Equilibria," *Journal of Economic Theory*, 1977, *14*, 32–44.

Lintner, John, "Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 1965, 13–37.

Manne, H. G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, 1965, *73*, 110–20.

Marris, Robin, *The Economic Theory of Managerial Capitalism*. Illinois: Free Press of Glencoe, 1964.

Milgrom, Paul, and Nancy Stokey, "Information, Trade and Common Knowledge," *Journal of Economic Theory*, 1982, *26*, 17–27.

Mirrlees, J. A., "The Theory of Moral Hazard and Unobservable Behavior—Part I," mimeo, Nuffield College, Oxford, 1975.

Radner, R., "Existence of Equilibrium in Plans, Prices, and Price Expectations in a Sequence of Markets," *Econometrica*, 1972, *40*, 289–303.

Radner, R., "Rational Expectations Equilibrium: Generic Existence and the Information Revealed by Price," *Econometrica*, 1979, *47*, 655–678.

Shleifer, Andrei, and Robert W. Vishny, "Large Shareholders and Corporate Control," *Journal of Political Economy*, 1986, *94*, 461–88.

Sharpe, William, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 1964, *19*, 425–42.

Stiglitz, Joseph E., "Information and Capital Market," mimeo, 1971.

Williamson, Oliver E., *The Economic Institutions of Capitalism*. New York: Free Press, 1985.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 137–148

A Seller's (& Buyer's) Guide to the Job Market for Beginning Academic Economists

Richard Carson and Peter Navarro

In a profession that traffics primarily in market processes and outcomes, it is both surprising and ironic that economists have not closely examined their own hiring process. No study has looked at the academic job market for new Ph.D. economists in any close institutional detail.[1]

This paper reports the major findings of a buyer's survey of the 1985–86 academic market for beginning Ph.D. economists based on a stratified random sample of all economics departments ranked in the top 20 (ECON20)[2] and 380 other economics departments (ECONOTHER).[3] Its purpose is to increase the level of

---

[1] In the past, Ernst Stromsdorfer has circulated a questionnaire to members of the AEA chairpersons' group which has focused on some aspects of the market outcome, but our survey focuses primarily on the process itself.

[2] Our ECON20 sample was determined by taking a composite of recent rankings (e.g., Laband, 1985). The departments included: Chicago, Columbia, Cornell, Harvard, Johns Hopkins, MIT, Michigan, Minnesota, NYU, Northwestern, Pennsylvania, Princeton, Rochester, Rutgers, Stanford, UC-Berkeley, UC-Los Angeles, UC-San Diego, Wisconsin, Yale. Our use of 20 schools was somewhat, but not totally, arbitrary: the data indicated that defining an ECON25 or perhaps an ECON30 group would not have significantly changed the results, but the data did not support going beyond that point.

[3] Our sample was drawn from 992 economics departments listed in the 1985 American Economics Association Handbook. All economics departments ranked in the ECON20 stratum were sampled and, free rider problems notwithstanding, all responded. 380 departments in the ECONOTHER stratum were sent surveys and 53 percent of those responded. All but one of the ECON20 departments had one or more job openings while a much lower percent (51 percent) of the 20 ECONOTHER departments which responded had one or more job openings.

■ *Richard Carson is Assistant Professor of Economics, University of California, San Diego and Peter Navarro is Assistant Professor of Economics, University of San Diego, both in San Diego, California.*

information available to buyers and particularly sellers in this market. Sellers in this job market are typically graduate students in the final stages of their doctoral dissertations and represent the latest vintage of economists graduating from the 115 universities that grant a Ph.D. They will find answers to questions like: Will a phone call from a candidate or faculty advisor increase the probability of securing a job interview? How many weeks before the AEA meetings are requests for interviews sent out? How long does a typical job interview last and what criteria are applied? How soon after the meetings interview is a candidate likely to be invited to give a seminar at a school? Are elements of the job offer such as salary, teaching load, and summer research money negotiable? How long does a candidate have to accept or reject an offer?

The benefits of these survey results will not, however, be limited to sellers. Buyers in this job market include economics and agricultural economic departments, business and public policy schools, and a variety of non-academic organizations ranging from government entities like the Federal Reserve and think tanks like the Brookings Institution to large corporations and small private consulting firms. For academic departments, this survey should provide a basis for comparing institutional practices. For those non-academic enterprises looking to hire Ph.D. economists, this survey will tell something about how the competition is hiring.

The Primary, Preemptive and Secondary Markets

The academic market for beginning Ph.D. economists may be divided into three segments.

The *primary market* or "meetings market," which will be examined in detail below, processed 75 percent of all the job openings represented by our survey. Job announcements in the fall issues of the American Economic Association's (AEA) *Job Openings for Economists* (*JOE*) traditionally mark the formal beginning of the recruitment season. On the basis of candidate applications and an important faculty information and intelligence network, buyers then typically choose a set of candidates to interview at the annual AEA meetings, site of perhaps the most intense exchange of information between buyers and sellers. In the next phase, buyers invite those candidates who have made the "second cut" to visit their schools or institutions and present their work during the first two to three months of the new year. For the successful candidate, a job offer may be extended during the visit or, more typically, with a lag of some weeks (and sometimes months). By April, the process is largely complete.

The *preemptive market* compresses interviews, campus visits, and offers into the time period before the December AEA meetings. We term this market preemptive because, based on our survey results and supplementary phone interviews with a number of department chairmen, this market appears to be largely limited to recruitment of the most sought after candidates in the market before the meetings market "auction" begins in earnest, particularly by the ECON20 departments. Consider, for example, the fact that 70 percent of the ECON20 departments participated in this market and accounted for 41 of the 82 campus visits.[4] Moreover, both the offer and rejection rates were quite high (and much higher than those rates in the meetings market reported below): *all* of the visits resulted in offers while 74 of the 82 offers were rejected. (One of the acceptances was at an ECON20 department.)

The *secondary market*, which accounted for 22 percent of all jobs filled in our sample, begins in January and extends into late May. For statistical purposes, we have defined this market to include only those job candidates who visited campuses after the AEA meetings but who were not interviewed at the meetings. It is a secondary market in the sense that it only starts to function after the primary meetings market has begun to clear. For example, 75 percent of the job candidates in this market were sought by departments who tried unsuccessfully to fill their openings through the primary meetings market or who posted their job openings in the *JOE* after the meetings. At the same time, none of the ECON20 departments participated in this market while the rejection rate for offers was, at 41 percent, substantially lower than the rate in either the preemptive or primary market. Finally, while the primary meetings market was largely over by the end of March, 41 percent of the secondary market offers were made in April and May.

Major Survey Results

Of the 223 economics departments in our sample, 54 percent were hiring and sought to fill 202 tenure track and 47 non-tenure track positions. Of these positions, 59 were at ECON20 departments (all tenure track). Advertisements appeared in the *JOE* for 93 percent of the positions, with 56 percent of the departments with job openings advertising in the October issue and 46 percent in the November issue. In addition, 39 percent advertised in the December issue, but the majority of these announcements were repeated from the October or November issues. Together, these three fall issues accounted for 77 percent of all the jobs advertised, while 55 percent of the departments advertised in more than one issue and only 48 percent advertised more than once during the fall. Besides using the *JOE* as a source of job information, 41 percent of all departments that sought to hire reported sending job announcements to an average of 58 other departments.

The departments seeking to hire reported 15,455 applications for tenure track and 1,498 for non-tenure track positions. The median ECON20 department reported receiving 150 applications per tenure track opening while the median ECONOTHER department reported 83 and 14 applications per tenured and non-tenured opening, respectively.

---

[4] This market may not be entirely preemptive. Fourteen percent of the ECONOTHER departments participated in this market, and roughly one-third of the visits to ECONOTHER departments appeared aimed at filling teaching needs for the spring semester or quarter.

Candidate applications were not, however, the only source of information used by departments to identify candidates. Indeed, while the median ECONOTHER department reported that 80 percent of its candidates were first identified by applications, the median ECON20 department had a corresponding figure of only 10 percent. What sources did the ECON20 departments rely on? In a statistic that both underscores the importance of an informal faculty "information/intelligence" network[5] in recruitment and illustrates a striking difference between the ECON20 and ECONOTHER departments, 50 percent of the candidates first identified by the median ECON20 department were recommended by a faculty member from the candidate's own department while 25 percent were identified by a faculty member within the recruiting department.[6]

The Meetings Market Interview Decision

Table 1 rates the importance of various aspects of the candidate's portfolio in the departmental decision to schedule an interview at the AEA meetings. In a convention used throughout the text in presenting tables, each cell has two percentage figures separated by a slash mark. The first represents the ECONOTHER departments and the second represents the ECON20 departments. Both numbers are presented because hiring practices clearly differ in many cases. To test these differences formally, a series of chi square tests of independence were conducted;[7] as a second convention, an asterisk placed by a category indicates that the hypothesis that the ECON20 departments differ from the ECONOTHER departments could not be rejected at the .05 level.

Table 1 illustrates that the candidate's research output together with faculty marketing inputs from the candidate's department such as phone calls and letters were very important in the departmental decision to interview while the candidate's field of specialization and such signals as school affiliation and years taken to complete the Ph.D. likewise were of some importance. With respect to faculty marketing inputs, letters of recommendation received the highest score of any factor, with 72 percent of the ECONOTHER departments and 82 percent of the ECON20 respondents rating them of great importance. At the same time, over half of the ECONOTHER

---

[5] After further inquiry about this network, we found that many departments, particularly the ECON20, formally or informally rank their own candidates and often pass this information on to departments that are hiring.

[6] "Resume books" sent by the candidate's department were also somewhat important: the median ECONOTHER and ECON20 departments identified 9 percent and 12 percent of candidates in this manner, respectively.

[7] The tests were performed on the 2 by X cross-tabulations formed by two types of departments and the X possible responses: for example, $X = 5$ in Table 1.

Table 1

The departmental decision to schedule an interview

	Great importance	*Moderate importance*	*Slight importance*	*No importance*	*Did not consider*
Candidate's letter of application*	11%/0%**	21%/6%	53%/18%	12%/63%	2%/13%
Enclosed unpublished papers*	25%/82%	38%/12%	30%/0%	4%/8%	4%/6%
Field(s) of specialization*	61%/24%	33%/41%	6%/29%	0%/6%	0%/0%
High probability thesis would be finished by start of job	42%/29%	36%/41%	15%/24%	5%/6%	2%/0%
Letters of recommendation	72%/82%	23%/18%	5%/0%	0%/0%	0%/0%
Number of years to finish doctorate	15%/6%	28%/29%	35%/53%	16%/12%	6%/0%
Phone call recommendations*	30%/47%	25%/41%	21%/12%	4%/0%	20%/0%
Phone call(s) from candidate expressing strong interest in your institution*	0%/0%	11%/12%	39%/12%	28%/53%	23%/24%
Previously published materials	51%/65%	40%/18%	9%/12%	0%/6%	1%/0%
Previous teaching experience*	32%/0%	36%/35%	25%/53%	6%/12%	1%/0%
Previous work experience	4%/0%	30%/12%	40%/59%	20%/24%	6%/6%
Resume*	42%/12%	34%/24%	20%/59%	5%/6%	0%/0%
School affiliation	39%/29%	51%/65%	10%/0%	0%/6%	0%/0%
School transcript	11%/0%	40%/24%	16%/18%	18%/41%	15%/18%
Thesis advisor(s) reputation*	25%/24%	51%/76%	19%/0%	3%/0%	3%/0%
Thesis chapter(s) quality*	34%/76%	35%/24%	19%/0%	5%/0%	8%/0%
Thesis topic	10%/18%	50%/65%	33%/18%	5%/0%	3%/0%
U.S. citizenship*	8%/0%	18%/0%	30%/18%	28%/53%	18%/29%

The first percent represents that of ECONOTHER departments while the second percent (after the slash) represents the ECON20 departments. An asterisk beside a category indicates that the hypothesis that ECON20 departments differ from the ECONOTHER departments could not be rejected at the .05 level for that category. Rows may not sum to 100 percent due to rounding.

departments and over 80 percent of the ECON20 departments rated the advisor's reputation and phone call recommendations as greatly to moderately important.

Note, however, that "candidate marketing inputs" such as the letter of application and candidate phone calls in support of an application appeared to be of little value: 67 percent of the ECONOTHER departments rated the letter of application as slight to no importance (or did not consider) as did 94 percent of the ECON20 respondents, while an even greater percentage rated the candidate phone call in similar fashion (90 percent and 89 percent, respectively).

Regarding research output, over 60 percent of the ECONOTHER departments and 70 percent or more of the ECON20 departments considered enclosed unpublished and previously published papers, the quality of the thesis chapter(s), and the probability that the thesis would be finished to be of great to moderate importance. However, the ECON20 chapters put an especially heavy emphasis on the unpublished papers, which are usually one or more thesis chapters. Eighty-two percent of the ECON20 departments rated such unpublished papers of great importance, while only 25 percent of the ECONOTHER departments did so.

In contrast, the candidate's graduate course work performance as summarized in the school transcript had relatively little bearing on the decision to interview: 49 percent of the ECONOTHER departments and 77 percent of the ECON20 respondents rated it as slight to no importance (or did not consider).[8] Teaching experience seemed to help candidates for ECONOTHER positions, although not with ECON20 departments, while work experience seemed to be of little assistance in obtaining an interview.

Finally, both school affiliation and, to a lesser extent, the number of years to complete the Ph.D. have some signaling value: over 90 percent of all departments rated affiliation as greatly or moderately important, while over 34 percent gave the number of years a similar rating.

The Meetings Market Interview

The median department in this survey invited 25 candidates to interview, the interquartile range was 18 to 35 candidates, and 10 percent invited over 40 candidates with a maximum of 55. These invitations were rarely refused. The median department began to schedule its interviews four weeks before the meetings, with 10 percent of all departments scheduling as early as six to seven weeks before and another 10 percent waiting until two weeks or less before the meetings. Interestingly, the ECONOTHER departments only scheduled 10 percent of their interviews *at* the meetings while for the ECON20 departments, the percent was only 3 percent. Thus, the probability of securing interviews at the meetings is quite small.

---

[8] Grades and performance in the classroom may, however, have an important indirect influence on faculty letters of recommendation and a department's consensus on who its top candidates are.

Table 2

Departmental evaluative criteria applied during the meetings interview

	Great importance	*Moderate importance*	*Slight importance*	*No importance*	*Did not consider*
Ability to ask incisive questions about your institution*	10%/0%**	48%/12%	30%/65%	9%/24%	3%/0%
Ability to explain thesis	66%/82%	24%/18%	9%/0%	0%/0%	1%/0%
Ability to respond to questions	78%/82%	21%/18%	0%/0%	0%/0%	1%/0%
Courses the candidate was willing/able to teach	32%/12%	52%/41%	15%/47%	0%/0%	1%/0%
Expressed strong interest in your institution*	15%/6%	59%/25%	18%/44%	6%/13%	1%/13%
Personableness*	19%/0%	68%/35%	11%/65%	1%/0%	1%/0%
Personal appearance*	5%/0%	47%/6%	34%/47%	10%/29%	3%/18%
Punctuality	6%/0%	30%/12%	42%/41%	14%/35%	8%/12%
Quality of future research agenda	56%/53%	32%/47%	9%/0%	3%/0%	1%/0%

The first percent represents that of ECONOTHER departments while the second percent (after the slash) represents the ECON20 departments. An asterisk beside a category indicates that the hypothesis that the ECON20 departments differ from the ECONOTHER departments could not be rejected at the .05 level for that category. Rows may not sum to 100 percent due to rounding.

The most common interview length was 30 minutes (55 percent of all departments); the next most common was 45 minutes (25 percent) while a few departments held interviews as short as 15 minutes and as long as 60 minutes. Two to three faculty members were typically present, although 10 percent of the departments reported having one member and 5 percent reported as many as four to five.

Table 2 rates the importance that the departments assigned to various aspects of the candidate's interview performance. Aspects of the candidate's research capabilities appear to play an important role together with the range of courses the candidate is willing or able to teach. For example, over 85 percent of the ECONOTHER and 100 percent of the ECON20 departments rated the candidate's ability to explain the thesis and respond to questions and the quality of the candidate's future research agenda to be of great to moderate significance while 84 percent of the ECONOTHER departments and 53 percent of the ECON20 gave the candidate's teaching portfolio a similar rating.

The ability of the candidate to ask incisive questions about the department's institution was of little importance at ECON20 institutions: 89 percent rated this factor of slight to no importance. Interestingly, however, 58 percent of the ECONOTHER departments found a candidate's demonstration of interest in the school to be of great to moderate importance in evaluating a candidate positively.

Finally, the candidate's personableness and personal appearance, if not punctuality, seemed to have great to moderate importance to over half of the ECONOTHER departments. Again, however, the ECON20 respondents significantly differed with over 60 percent finding personableness and personal appearance to be of only slight to no importance (or did not consider).

The Campus Visit

From the pool of candidates interviewed at the meetings, the median department invited 5 candidates for a campus visit, the interquartile range was 3 to 8 candidates, and 10 percent of the departments invited 12 or more to a maximum of 23 candidates. Only a small percentage of campus visit invitations were turned down. Visits typically began the second week of January and were largely concluded by the end of February, although some visits occurred as late as May. Virtually all departments reimbursed candidates for travel expenses.

Table 3 rates the importance that departments assigned to 11 different aspects of the candidate's performance during the visit. The candidate's research again plays a major role in the evaluation process: 81 percent of the ECONOTHER and 84 percent of the ECON20 departments rated the overall seminar presentation to be of great importance, while over 90 percent of all departments rated the ability to answer questions, the clarity of the seminar presentation, and the ability to present technically difficult subject matter to be of great to moderate importance. The only other criterion that appears to approach these aspects of the seminar in importance was faculty visits with the candidate: over 85 percent of all departments found them to be of great to moderate importance.

Visits with administration officials were somewhat important at the ECONOTHER schools but not at the ECON20 schools: 51 percent of the ECONOTHER departments rated candidate visits with administration officials greatly or moderately important as opposed to only 11 percent of the ECON20. In addition, the smaller the institution and the greater its teaching versus research orientation, the more important meetings with administration officials appear to be. Faring the worst was visits with students: 61 percent of the ECONOTHER departments and 89 percent of the ECON20 respondents rated these visits to be of slight to no importance (or did not consider).

As a final comment, two clear differences between the hiring practices of the ECON20 and the ECONOTHER departments that have already emerged are

Table 3

Evaluative criteria for the campus visit

	Great importance	*Moderate importance*	*Slight importance*	*No importance*	*Did not consider*
Demonstrated social skills*	12%/0%**	50%/11%	33%/61%	4%/28%	1%/0%
Individual visits with faculty*	55%/22%	43%/67%	2%/6%	0%/6%	0%/0%
Interest expressed for your school*	18%/6%	59%/39%	20%/33%	2%/17%	0%/6%
Visiting with students*	11%/0%	27%/11%	24%/33%	13%/39%	24%/17%
Visits with administration officials (e.g., Dean)*	16%/0%	35%/11%'	30%/11%	14%/39%	4%/39%
SEMINAR ASPECTS					
Overall seminar presentation	81%/84%	16%/16%	1%/0%	0%/0%	1%/0%
Ability to answer questions	76%/84%	22%/16%	1%/0%	0%/0%	1%/0%
Ability to present technically difficult subject matter	51%/50%	43%/44%	5%/6%	0%/0%	1%/0%
Clarity of presentation	77%/78%	22%/22%	0%/0%	0%/0%	1%/0%
Demonstrated teaching ability	59%/17%	26%/50%	11%/33%	1%/0%	3%/0%
Knowledge of contributions in related fields	22%/17%	52%/39%	21%/44%	2%/0%	2%/0%

The first percent represents that of ECONOTHER departments while the second percent (after the slash) represents the ECON20 departments. An asterisk beside a category indicates that the hypothesis that the ECON20 departments differ from the ECONOTHER departments could not be rejected at the .05 level for that category. Rows may not sum to 100 percent due to rounding.

reinforced here. First, teaching ability again appears to be of less importance to the ECON20 versus ECONOTHER departments, as evidenced by the 17 percent versus 59 percent great importance rating attached to this aspect of the seminar presentation. Second, the expression of interest by a candidate in an institution has less signaling value to the ECON20 departments than the ECONOTHER departments, as evidenced by the 45 percent versus 77 percent moderate to great importance rating assigned to this factor.

The Job Offer Process

The ECONOTHER departments reported making 174 job offers to candidates participating in the meetings market while the ECON20 departments made 88 offers. The median department for the combined sample invited 5 candidates to give seminars and extended job offers to 2 of them, the interquartile range for the number of offers extended was 1 to 4 offers, and 10 percent of the departments made 5 or more offers with the maximum being 12. Interestingly, slightly more than 10 percent of the departments made no offers despite inviting one or more candidates to visit.

Job offers come as early as January, with heavy action in February; and the market is largely over by the end of March. In examining the pattern of offers, there does appear to be some evidence that the ECON20 market clears earlier than the ECONOTHER market: 41 percent of the ECON20 offers were made in January as opposed to 16 percent for the ECONOTHER, while 98 percent of all ECON20 offers were made by the end of February as opposed to 70 percent for the ECONOTHER departments.

Sixty-five percent of all departments gave the typical candidate between 10 to 14 days to accept their offer, while the median time was 14 days. There was, however, apparent leeway in this consideration time: 10 percent of departments reported that the longest time was 30 days or more, with one department reporting a time of 180 days while 10 percent reported that the shortest time was 4 days or less with a few cases of requiring an immediate response. Candidates accepted 47 percent of the ECONOTHER offers and a somewhat lower 37 percent of the ECON20 offers. Within the ECON20 departments, there was a noticeable increase in the acceptance rate as one moved up the rankings ladder. However, every ECON20 department seeking to hire experienced at least one rejection.

Summing job offers accepted over the three markets (that is, primary, preemptive, and secondary) and comparing this figure to the total jobs departments sought to fill reveals that a significant percentage of jobs went unfilled in our sample: 31 percent in the ECONOTHER departments and 44 percent in the ECON20 departments. This pattern is consistent with statistics reported for previous years.

Characteristics of the Job Offer

Most salary offers were clustered in a very tight range for all departments, although the ECON20 salary offers were consistently $2000 to $3000 higher across the range and the ECONOTHER offers exhibited slightly more variability.

For example, the median ECON20 department's average offer was $32,500 while the interquartile range was $32,000 to $34,500; there were only two offers above $36,000 and only two offers below $31,000. Similarly, the median ECONOTHER departments reported an average offer of $30,000 while the interquartile range was $28,000 to $31,500 with only four offers above $36,000 and four below $24,000.

In extending these offers, 60 percent of the ECON20 departments reported substantial discretion and 15 percent reported total discretion in matching the salary offers of competing economics departments while the corresponding figures for the ECONOTHER departments were somewhat lower at 39 percent and 9 percent, respectively. This widespread ability to match competing offers from other economics departments may well explain why salary offers vary around such a narrow band. This discretion did not extend, however, to meeting offers from business schools or non-academic institutions: only 32 percent of the ECON20 and 24 percent of the ECONOTHER departments reported substantial discretion in meeting these offers while none of the ECON20 and only 6 percent of the ECONOTHER departments reported total discretion.

Besides offering slightly higher salaries, over 90 percent of the ECON20 departments reported that summer research money was either usually available or a negotiable item as opposed to only 43 percent of the ECONOTHER departments. Because summer money is often equivalent to at least one-ninth of regular salary, this suggests the *effective* salary gap between the ECON20 and ECONOTHER departments may be larger than it appears.

A similar pattern of differences prevailed for a new hire's course and preparation loads which were only slightly lower than for regular faculty: the ECONOTHER department interquartile range was four to six courses and three to four preps while the ECON20 department range was three to four courses and two to three preps. Only 6 percent of those accepting jobs successfully bargained for a reduced teaching load. Other significant negotiated aspects of the job compensation package included choice of classes to teach and, to a lesser extent, availability of research assistance and computers. Subsidized housing was almost never negotiated.

Summary and Conclusions

In an effort to increase the stock of information available to sellers and buyers in the academic job market for beginning Ph.D. economists, this paper has presented the findings of a survey of the 1985–86 hiring process by economics departments.

1. There are three sub-markets operating within the overall job market. The "meetings market" sorts and clears rather quickly: over 90 percent of all job offers in this market are made in January and February. The ECON20 departments are major players in a "preemptive market" that occurs before the "meetings market" begins and is aimed primarily at the very top candidates in the market. In contrast, the ECON20 departments do not participate in a "secondary market" which primarily sorts during and after the clearing of the meetings market.

2. The October and November issues of the *JOE* are the most important sources of information about job openings while a high percentage of advertised jobs (over 30 percent) go unfilled each year.

3. An important faculty information/intelligence network helps to identify prospective job candidates, particularly among the ECON20 departments. For sellers

in this market, this means that faculty letters, phone calls of recommendation, and being rated as one of their department's top candidates can be crucial.

4. The job candidate's unpublished and published papers, the quality of his or her seminar presentation, and the future research agenda are the most important keys to success in the job market.

5. The hiring practices of the ECON20 and ECONOTHER departments differ substantially. In general, ECON20 departments rely more on the faculty information/intelligence network to find candidates, put a higher value on research skills and a lower value on teaching skills, offer slightly higher base salaries and more "fringes," and require a lower teaching load.

6. Salaries vary around a narrow band. Whether this is due to cartel behavior among departments or whether the observation of "one price" indicates a highly competitive market is, however, difficult to determine. Most departments have substantial discretion in matching offers by competing departments and there is considerable variation in non-base salary compensation and perquisites in addition to quality competition.

7. Based on the annual number of Ph.D. economics graduates in previous years, we estimate that slightly less than 25 percent get tenure track positions in American four-year colleges and universities while slightly less than 5 percent get jobs at ECON20 departments. There is also a substantial flow of new Ph.D. economists to business schools and, to a lesser extent, other professional schools and agricultural economics departments, while a large number of graduates find employment in the private and public sectors.

■ *The authors would like to thank the University of San Diego for generous financial support.*

References

Laband, D.N., "An Evaluation of 50 'Ranked' Economics Departments—By Quantity and Quality of Faculty Publications and Graduate Student Placement and Research Success," *Southern Economic Journal*, 1985, *52*, 216–240.

The College Blue Book. 20th edition. New York: MacMillan, 1985.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 149–156

Puzzles
Cider in Your Ear, Continuing Dilemma, The Last Shall Be First, and More

Barry Nalebuff

In presenting economic puzzles, I have three goals in mind: some puzzles are chosen to stimulate research; others offer examples that will help undergraduate and graduate teaching; all should provide quality distractions during seminars. As usual, this feature begins with several speed puzzles; answers can be found at the end of the problems. Following are several longer puzzles, for which readers are invited — nay, challenged, to submit their own answers. The puzzles in this issue focus on betting and voting, and once again include a prisoners' dilemma problem. The column ends with reader mail, including the submitted solutions to "The Best Location in Manhattan," Puzzle 4 in the Summer 1987 issue of this journal.

As always, please send your answers and favorite puzzles to: Barry Nalebuff, "Puzzles," c/o *Journal of Economic Perspectives*, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton NJ 08544. Good luck.

Puzzle 1: Here's Cider in Your Ear

Consider the following excerpt from Huff and Geis (1959) *How to Take a Chance*.

> Says Joe then: "It's a good night to play matching pennies—too hot for anything more strenuous. In fact, it's pretty warm to flip actual coins, so let's just lean back and *say* 'Heads' or 'Tails' instead." [*Assume these are called out simultaneously.*]
>
> "All right so far," you agree.
>
> "And to put a little variety into the game," Joe goes on, "I'll give you $3 when I call tails and you call heads. I'll give $1 when it is the other way around

■ *Barry Nalebuff is Assistant Professor of Economics, Princeton University, Princeton, New Jersey.*

> (i.e., I call heads and you call tails). And when we match, you give me $2 to make it even."
>
> At this point you do a little figuring. On the basis of pure chance, which would apply if this were actual coin tossing, the offer is fair enough. In the long run, out of each four tosses you would win $3 once and $1 once and lose $2 twice, for an even break.

If you play against Joe, what do you expect to earn (and how would you play)? So would you play?

Puzzle 2: When You Can't Beat Them One-on-One

A serious problem with voting under 50 percent majority rule is that there may fail to exist a proposal which beats all other proposals in pairwise competition. The standard Condorcet cycle is but one of many examples: if there are three individuals with strict preferences $A > B > C$; $B > C > A$; and $C > A > B$, then in pairwise votes A beats B, B beats C, and C beats A.

When no pairwise victor emerges, a possible alternative might be to choose a proposal (for example, A) that beats all other proposals in pairwise competition, either directly (as in A beats B) or indirectly in *one* step (as in A beats B and B beats C so A indirectly beats C in one step).

Does such a proposal always exist?

Puzzle 3: The Dilemma Continues

Abraham Neyman offers the following twist on the prisoners' dilemma. This provides a good test for your understanding of game theory. It is well known that there is a unique subgame perfect equilibrium in the finitely repeated prisoners' dilemma. It is also true that there is a unique Nash equilibrium in the finitely repeated prisoners' dilemma. How is this argument different from that of subgame perfection?

Answers to Puzzles 1, 2, and 3 appear following Puzzle 5.

Puzzle 4: The Other Person's Envelope is Always Greener

You have two envelopes. In one, you place a hidden amount of money and give the envelope to Ali. Then you flip a hidden coin. If it comes up heads, you place twice the original amount of money in the second envelope. If it comes up tails you only put half the original in the second envelope. You give this second envelope to Baba. So far, the contents of both envelopes are hidden, as is the outcome of the coin toss. Ali and Baba are allowed to look privately at the amount of money in their own envelopes. They are then given an opportunity to trade envelopes if they both agree.

Suppose, for the sake of argument, that Ali finds $10.00 in her envelope. Ali reasons that Baba is equally likely to have $5.00 or $20.00. Trading envelopes gives her an expected gain of $2.50 (or 25 percent). Acting in a risk-neutral manner, she would want to switch.

Now Baba looks inside his envelope. Whatever amount he finds (either $5.00 or $20.00), he too reasons that Ali is equally likely to have half or double his amount. The expectation is $0.5[0.5X + 2X] = 1.25X$, so he too expects a 25 percent gain from switching envelopes.

But this is paradoxical. The sum of the amounts in the envelopes is whatever it is. Trading envelopes simply cannot make *both* students better off. Yet they both expect to make a 25 percent gain. Where did they go wrong?

I learned about this puzzle from Hal Varian who heard it from S. Zabell at Northwestern who heard it from Steve Budrys of the Odesta Corporation. Although it may not have originated with Budrys, this is as far back as we have been able to trace it. I would be delighted to hear more about the history of this puzzle. My telling above involves a slight variation on the "original" version, where there is no coin toss. We are simply told that one envelope is twice as nice as the other, but not which is which. So before the envelopes are handed out, the two students should be indifferent as to which they get. But once they open their envelopes, each appears to be eager to trade with the other. Just as above, both expect a 25 percent return from the trade. In fact, even before they look in their envelope, they both want to trade. How can that be? For those readers who are impatient, Zabell (1987) provides a short exposition of the puzzle and his proposed solution. I will discuss his and your solutions in a later column.

Puzzle 5: The Last Shall be First?

One of my all-time favorite puzzles is to figure out what to do in Tom Schelling's paradox of the departing voter below.

Schelling describes the set-up: "A five-man board is to elect one of its members chairman by a procedure involving successive majority votes. Anderson, first in alphabetical order, will be paired against Barnes; the winner of that vote will be paired against Carlson, the winner then paired against Davis, and the winner of that paired against Evans. The winner of this fourth and final ballot will be declared chairman.

Everyone knows everyone else's preferences. Everyone wants to be chairman. Reading (initial letter only) left to right in the table below are the ways that each of the five members ranks all five as candidates for the job. Anderson, for example, ranks them all in alphabetical order. Barnes, for example, ranks Evans third and Carlson last.

Anderson	A	B	C	D	E
Barnes	B	A	E	D	C
Carlson	C	D	A	E	B
Davis	D	B	A	E	C
Evans	E	D	B	C	A

All five committee members have foresight and vote strategically. Who will win? So far, the answer is straightforward, but I'll let you figure it out.

Now for the Schelling fillip. Anderson is forced to miss the meeting. He is allowed to transfer his vote to someone else. This transfer must be unconditional in that Anderson is not allowed to specify how the receiver must vote. To whom should Anderson give his vote and what do you expect will happen?

Answers to Speed Puzzles

Answer to Puzzle 1

In *Guys and Dolls*, gambler Sky Masterson relates this valuable advice from his father: "Son, one of these days in your travels a guy is going to come to you and show you a nice brand-new deck of cards on which the seal is not yet broken, and this guy is going to offer to bet you that he can make the jack of spades jump out of the deck and squirt cider in your ear. But son, do not bet this man, for as surely as you stand there you are going to end up with cider in your ear."

Joe has just offered you one of those cider in your ear bets. If he were to randomize 50–50, as with a fair coin, then you would expect to break even. But Joe can do even better by playing a mixed strategy, calling heads with chance 5/8 and tails 3/8. Whether you call Heads or Tails, Joe has guaranteed an expected payoff of 1/8; you have cider in your ear (and an expected $-1/8$ payoff) if you accepted his bet.

Answer to Puzzle 2

This question was first answered by Landau (1953) in the context of determining hierarchies in animal colonies. Think of each proposal as one chicken in a flock. In pairwise competition, the two chickens are ranked based on who can peck the other. Landau demonstrated the existence of a Chicken King in the sense that this chicken can beat (up) every other chicken either directly or via some third chicken. In particular, the chicken which beats up the greatest number of other chickens has this property. The proof is straightforward. Call the chicken which beats the most other chickens K. If K beats all other chickens then we are done. If K loses to some chicken L then it must be the case that K beats something which beats L. Otherwise, L would beat everything K beats and K too: this would give L more victories than K, contradicting the assumption that no chicken has more victories than K.

Returning to the question of ranking social alternatives, Nicholas Miller (1980) offers an elegant solution in his article on the uncovered set. Miller defines the "Cover" relationship by *A* covers *B* if (1) *A* beats *B* in a pairwise majority vote and (2) *A* also beats everything that *B* beats (in pairwise majority votes). Miller shows that this covering relationship is transitive; if *A* covers *B* and *B* covers *C* this implies that *A* covers *C*. Since a transitive relationship cannot cycle in any finite collection of alternatives, there must be at least one "Uncovered" proposition. I leave it to the reader to show that a proposal uncovered by anything else has the "Chicken King" property, namely it beats everything else either directly or in one step.

The importance of the uncovered set goes beyond the existence of a "Chicken King." As emphasized by Ordershook (1986) it allows us to restrict the problem of social choice to the set of uncovered proposals; why pick *B* if *A* covers *B*? Although this may not reduce the choice to a singleton, it is a stronger selection criteria than the Pareto principle alone. Second, Miller (1980) demonstrates that sophisticated voting under *amendment procedure* can never lead to an outcome outside the uncovered set. Hence, the ability to manipulate an agenda is restricted to the set of alternatives that are uncovered.

Answer to Puzzle 3

First, the subgame perfection argument. In the final period, there is a unique Nash equilibrium: both players defect. Hence behavior in the penultimate period cannot affect the final play, so again both players should defect. And so it goes.

The Nash argument is more subtle. Again, at the final period both players must defect in the Nash equilibrium outcome. But that does not immediately imply that in the penultimate period both players must defect in any Nash equilibrium outcome. If in the penultimate period, player 1 deviates from the predicted equilibrium behavior, player 2 is no longer required to behave rationally in final period. But in the prisoners' dilemma example, there is nothing worse that he can do than defect. Player 2's equilibrium strategy already *minimizes* player 1's final period payoff. There is no off-equilibrium punishment that can be used to generate cooperation in earlier periods.

This difference is easier to see in Table 1 where the prisoners' dilemma is expanded to include a third strategy, mutual destruction (confession to a crime that

Table 1
Prisoners Dilemma with Mutual Destruction

			Prisoner 2	
		Cooperate	Defect	Mutual Destruction
	Cooperate	1, 1	5, 0	−99, −99
Prisoner 1	Defect	0, 5	3, 3	−99, −99
	Mutual destruction	−99, −99	−99, −99	−99, −99

will give both prisoners the death penalty). Mutual destruction is clearly a dominated strategy and will never be played in any subgame perfect equilibrium. But in a Nash equilibrium, the prisoners could use this "incredible" threat of mutual destruction to sustain cooperation in all but the last period of a finite repetition of the game above. This is accomplished by the pair of equilibrium strategies: cooperate in period $t + 1$ if the other has cooperated in periods 1 to t, else mutual destruction.

Mail

Comments on Puzzles from Fall 1987

Puzzle 3. Two for the Team, but not Tea for Two. It turns out that not all college football teams made Nebraska's mistake. In the 1962 Yale–Colgate football game, the Elis were trailing 14–0 in the final quarter. After their first touchdown, they went for the two point conversion but failed. But all was not lost. They scored a second touchdown, made their second attempt at a two point conversion and tied the contest (See Porter, 1967).

Richard Porter (Michigan) is the expert on this topic and has a wonderful two-page article "Extra-Point Strategy in Football." His note is ideal for introducing revealed preference, expected utility, and dominated strategies to undergraduates. It is packed with fascinating examples from college football. As for those who haven't read Porter, they end up condemned to rewrite it, either by ignoring his conclusions (Nebraska) or rediscovering them (me). It's better in the original, both the sin and the solution.

Comments on Puzzles from Summer 1987

Answer to Puzzle 4: The Best Location in Manhattan. This puzzle asked for the most efficient pattern of firm locations when consumers are uniformly distributed over a two-dimensional plane, and consumers shop at the "closest" store, where closest is based on the Manhattan metric. Under the Manhattan metric, the distance between A and B equals the sum of the vertical and horizontal differences; if $A = (a_1, a_2)$ and $B = (b_1, b_2)$ then the distance between A and B, denoted by $\|A - B\|$, equals $|a_1 - b_1| + |a_2 - b_2|$. With this metric, what is the most efficient location pattern and what is the resulting shape of market areas?

No, the answer is not squares! But if you guessed squares, you are in good company. The correct answer is Diamonds. Along the border of a diamond, all customers travel exactly the same distance to the center. This shape minimizes the average (Manhattan) distance travelled for a given market area. For example, we see in Figure 1 that the Manhattan metric distance to any point on the border of a diamond is $\sqrt{2}\,Y/2$ when the market area per store is Y^2. In contrast, if the same density of firms were to locate in a square pattern, then wherever the square sticks out over the diamond (the four shaded corners) these customers will have to travel farther, all the way up to Y at the corners. The same "overlapping" argument demonstrates that over all market areas of Y^2 the diamond shape minimizes the average Manhattan distance travelled.

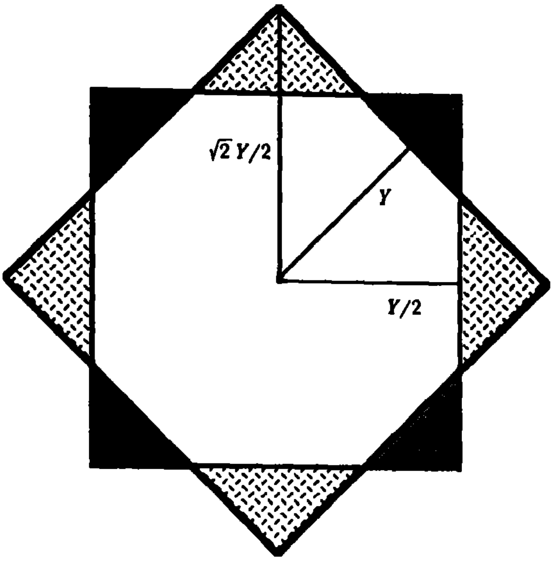


FIGURE 1

In order received, correct answers were submitted by John Brown (Albion College), R.D. Cairns (McGill), James Dow (Univ. of Penn.), and Curtis Eaton (Simon Fraser University).

But Curtis Eaton wins the prize for sending in a proof. One cannot just assume as above that the solution is symmetric; this has to be demonstrated. Curtis went on to write: "What about configurations that are not area-symmetric? If there are any non-diamond shaped markets, total and hence average distance from the nearest firm would decrease if we could 'diamond' them, leaving all the other markets the same [for the reason given above]. In general, we could not 'diamond' them, leaving all the other markets the same. But let's suppose that we could. Now consider any pair of diamond markets with unequal area. We could again reduce total and hence average distance if we could produce identical diamond markets for them (leaving all other markets the same), since total distance from a firm with a diamond market with side Y is $\sqrt{2}\,Y^3/3$, which is a convex function of Y. Again, we could not obviously produce the identical diamond market areas, but suppose that we could. Replication of the 'diamonds of unequal sizes' argument leads us in a series of imagined steps, each of which reduces average distance, to the conclusion that all diamond market areas must be identical. We can, of course, implement this configuration."

More on Answer to Puzzle 2: In Fact, It's a Gas. The controversy over the leaded-unleaded gasoline price differential continues. John Lott Jr. (Rice University) and Russell Roberts (UCLA) write to offer an alternative explanation of the price premium for unleaded gasoline that does not depend on differing price elasticities nor loss leader effects. They argue that the price differential is due to gasoline stations attempting to cover their fixed costs such as checking the oil, turning on the pump, making change, and so on. Consumers of leaded gas typically have cars with bigger gasoline tanks since their cars are older and less fuel efficient. Since their purchases are bigger, the fixed cost can be covered with a smaller mark-up on the larger transaction. They draw an interesting analogy to price discrimination in restaurants. "The most important factor explaining the retail over wholesale mark-up in a restaurant is the time it takes to eat an item. This explains why wine and coffee are priced seemingly so far above 'marginal cost.' People linger over these items, so their price must include the cost of renting the table." While I find the restaurant analogy helpful, I think their explanation is not entirely convincing. If the markup on coffee is an attempt to cover the rental cost of a table, why then are refills typically free and why is there no discount for take out? As for gasoline, the fixed cost associated with self-service must be lower. This would imply a smaller price differential between grades at the unleaded pumps. Is that true? Lott and Roberts are aware of these issues and have promised to drive from coffee shop to gas station until they find the answer.

As for New Jersey, Burt Malkiel (Yale) recalls the legislative debate when self-service gasoline was outlawed. The "argument" was that self-service pumps discriminate against women.

More on Puzzle 1: Turn Out the Lights. Peter Albin (CUNY—John Jay College) asked if he "might shed additional darkness on Puzzle 1 before Consolidated Edison takes the matter as settled doctrine and mails out my priority vouchers." The problem with the solution as given is that it requires that brownouts have no significant rationing or signalling effects on their own. This might not be the case if in the event of a brownout "certain equipment is turned off entirely because it cannot function if plug current does not meet design specifications: e.g. computers without independent power supplies are switched off." Since the priority system might allow others to continue using their computers, this could exacerbate the total shortage. "Although the net value of transferred priority vouchers will always be zero, the status need not be quo." I plead guilty as charged.

■ *Behind the scenes Carl Shapiro and Timothy Taylor have been improving the puzzles and the answers. Thomas Schelling at Harvard's JFK School has been kind enough to share this paradox of the departing voter. The idea of an "uncovered set" is based on an article by Nicholas Miller (1980). Finally, many thanks to all of those writing in with answers and new puzzles.*

References

Huff, Darrell, and Irving Geis, *How to Take a Chance*, New York: Norton, 1959.

Landau, H. G., "On Dominance Relations and the Structure of Animal Societies:III. The Conditions for a Score Sequence," *Bulletin of Mathematical Biophysics*, 1953, *15*, 143–148.

Miller, Nicholas, "A New Solution Set for Tournaments and Majority Voting: Further Graph-Theoretical Approaches to the Theory of Voting," *American Journal of Political Science*, 1980, *24*, 68–96.

Ordeshook, Peter, *Game Theory and Political Theory*, Cambridge: University of Cambridge Press, 1986.

Porter, Richard C., "Extra-Point Strategy in Football," *The American Statistician*, 1967, *21*, 14–15.

Zabell, S. L., "Loss and Gain: The Budry's Paradox," mimeo, Dept. of Mathematics and Statistics, Northwestern University, 1987.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 157–160

Recommendations for Further Reading

Bernard Saffran

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, the articles chosen will be expository or integrative and not focus on original research. As a rule of thumb, articles listed in this section appear in journals that are not indexed by the *Journal of Economic Literature* (which focuses on the research journals) and are similar to many of the articles published in this journal. They may include survey articles, discussion of related subjects in which economists might have an interest (for example, the use of game theory in evolutionary biology) or analyses of economics from other perspectives.

The intention is to publish a selective list of ten to fifteen articles per issue, with the selections depending heavily on input from readers. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Bernard Saffran, c/o *Journal of Economic Perspectives*, Department of Economics, Swarthmore College, Swarthmore, Pennsylvania 19081.

Mixture and Medley

These articles range from the possibility of detrimental competition to the limits of international cooperation, from methodology in econometrics to methodology in accounting, from hot streaks in basketball to statistical mechanics, from business forecasting to appropriate technology and more.

Adams, F. Gerard, "Revolution in the World of Econometric Business Forecasting." *Business Horizons*, Indiana University School of Business, January/February 1987, 46–51. Based on his book, *The Business Forecasting Revolution*, Oxford University Press, 1986, the article discusses how the forecasting function has moved down the organizational ladder from the chief financial officer's office to the marketing,

■ *Bernard Saffran is Professor of Economics, Swarthmore College, Swarthmore, Pennsylvania.*

distribution and planning department. Professional economists now do less of the business forecasting and this has considerable risks.

Bell, Philip W., "Accounting as a Discipline for Study and Practice: 1986," *Contemporary Accounting Research*, Spring 1987, 338–367. A discussion of methodology in accounting. After surveying the developments of recent years he argues for the importance of normative standards to complement the work of the "Positive Accountants."

Eckaus, R. S., "Appropriate Technology: The Movement Has Only a Few Clothes On," *Issues in Science and Technology*, Winter 1987, 62–71. An overview of the accomplishments and pitfalls of the appropriate technology movement which grew out of E. F. Schumacher's criticisms in the late 1960s of the transfer of Western technologies to developing countries. "He faults the AT movement's reasoning, arguing that it is vague and based on technological determinism. And he claims that the technologies developed by AT organizations have contributed little to economic development." A response is given by Frances Stewart in "The Case for Appropriate Technology: A Reply to R. S. Eckaus," *Issues in Science and Technology*, Summer 1987, 101–109. "Eckaus does AT a disservice by characterizing it as a monolithic movement driven only by technological determinism. AT is instead a concept designed to indicate a broad direction of change, not a uniform doctrine for all countries."

Gilovich, Thomas, Robert Vallone, and Amos Tversky, "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, *17*, 1985, 295–314. A great article to use for a lecture on random sequences. Basketball players and fans believe that a player is more likely to make a basket if he has just made one. The authors find no evidence for this effect. "The belief in the hot hand and the 'detection' of streaks in random sequences is attributed to a general misconception of chance according to which even short random sequences are thought to be highly representative of their generating process."

Laidler, David, "'Buffer-Stock' Money and the Transmission Mechanism," *Economic Review*, Federal Reserve Bank of Atlanta, March/April 1987, 11–23. The November 1986 Atlanta Fed's Distinguished Lecture. Defense of the "buffer stock" money view against those of the orthodox Keynesian and new-Classical economics. He concludes that "the buffer-stock approach... predicts that systematic policies can indeed have real effects. However, it warns that their magnitude and timing are sufficiently uncertain as to render them positively dangerous."

Pagan, Adrian, "Three Econometric Methodologies: A Critical Appraisal," *Journal of Economic Surveys*, *1*, No. 1, 1987, 3–24. A critical comparison of the econometric methodologies proposed by David Hendry, Christopher Sims and Edward Leamer.

Horlick, Max, "The Relationships Between Public and Private Pension Schemes: An Introductory Overview," *Social Security Bulletin*, July 1987, 15–24. An examination of the growth and integration of public and private pensions in the U.S. and Europe and the possibility of using individual tax encouraged savings as a "third pillar."

Montroll, Elliot W., "On the Dynamics and Evolution of Some Sociotechnical Systems," *Bulletin* (New Series) *of the American Mathematical Society*, January 1987, 1–46. This Josiah Willard Gibbs Lecture by a student of statistical mechanics presents a fresh viewpoint on the use of certain elementary techniques. There is a discussion of the application of the techniques to prices, income distribution and technical change. He also has a fine sense for interesting tidbits; for example, I did not know that J. P. Morgan was offered a privatdozentship in mathematics at Göttingen University by Fritz Klein.

Schweitzer, Arthur, "Detrimental Competition," *Journal of Economic Issues*, September 1986, 681–707. A reformulation of the theory of competition. In detrimental competition the focus is on avoiding losses that arise from uncertainty and cutthroat competition. Instead of being universally beneficial, competition in markets with sustained surplus leads to distress prices that call for the principle of social minimum and for enforced minimum wages.

"The World Economy—Living with Uncertainty," *The Economist*, September 26, 1987, pull-out survey. A classical *Economist* survey by economics editor Clive Crook. He concludes, "International co-operation in pursuit of modest, right-minded goals cannot do much harm... But let them think of co-operation as a boring means to an end, not as a glorious goal in its own right. Because if they all stayed at home and adopted sensible domestic policies there would be precious little need for co-operation on trade or exchange rates."

Something Extra From *Science*

Science magazine continues to provide terse and readable articles on economics that can serve as updates or introductions to a subject.

Briggs, Vernon M., Jr., "The Growth and Composition of the U.S. Labor Force," *Science*, October 9, 1987, 176–180. "In sharp contrast with the experiences of all other industrialized nations, the size of the labor force of the United States is growing rapidly while, simultaneously, its age, gender and ethnic composition are changing markedly... New public policies that focus on labor market adjustment policies will be required if these developments are to be a boon rather than a bane to the emerging postindustrial economy."

Eisner, Robert, "The Federal Deficit: How Does it Matter?" *Science*, September 25, 1987, 1577–1582. A good, short, clear introduction to Eisner's well-known views. "Official measures of deficits are misleading, however, for failing to distinguish public investment from current spending and for failing to adjust for inflation. Correctly measured real deficits have contributed to the growth of gross national product, consumption and imports, and to investment as well."

Huber, Peter, "Injury Litigation and Liability Insurance Dynamics," *Science*, October 2, 1987, 31–36. "Prices for some lines of liability insurance have increased sharply in recent years, even while the real amount of coverage has declined... the most likely underlying cause for the current crisis in liability insurance is the inexorable expansion in liability law."

Personalities and the Profession

Two interesting symposia have appeared recently, and conversations continue to become ever more popular.

"Symposium on *Frank Knight*," *Ethics*, October 1987, 61–103. These articles by James M. Buchanan, Jules Coleman and Richard E. Flathman stem from a Conference on Knight on the occasion of his one hundredth birthdate.

"Development Economics After 40 Years. Essays in Honor of Peter Bauer," *Cato Journal*, Spring/Summer 1987, 1–254. The many distinguished authors emphasize the important contributions that Bauer has made to the changing attitude towards government intervention and markets in development economics.

"An Interview with Kenneth J. Arrow," by J. S. Kelly, *Social Choice and Welfare*, *4*, 1987, 43–62.

"A Conversation with C. R. Rao," by Morris H. DeGroot, *Statistical Science*, *2*, No. 1, 1987, 53–67.

■ *I would like to thank Anil K. Bera, David L. Cleeton, Ned Prescott, Peter Siegelman, Sherrill Shaffer and Timothy Taylor for their helpful suggestions.*

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 161–174

Anomalies
Parimutuel Betting Markets: Racetracks and Lotteries

Richard H. Thaler and William T. Ziemba

Economics can be distinguished from other social sciences by the belief that most (all?) behavior can be explained by assuming that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear. An empirical result qualifies as an anomaly if it is difficult to "rationalize," or if implausible assumptions are necessary to explain it within the paradigm. This column will present a series of such anomalies. Of course, "difficult" and "implausible" are judgments, and others might disagree with my assessment. Therefore, I invite readers to submit *brief* explanations (within the paradigm or otherwise) for any of the anomalies I report. To be considered for publication, however, proposed explanations must be falsifiable, at least in principle. Future topics for this column will come from as many fields of empirical economics as possible Readers are invited to suggest topics by sending a note with some references to (or better yet copies of) the relevant research. The address is: Richard Thaler, c/o *Journal of Economic Perspectives*, Johnson Graduate School of Management, Malott Hall, Cornell University, Ithaca, NY 14853.

Introduction

Economists have given great attention to stock markets in their efforts to test the concepts of market efficiency and rationality. Yet wagering markets are, in one key

■ *Richard H. Thaler is the Henrietta Johnson Louis Professor of Economics at the Johnson Graduate School of Management, Cornell University, Ithaca, New York. William T. Ziemba is Alumni Professor of Management Science, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada.*

respect, better suited for testing market efficiency and rationality. The advantage of wagering markets is that each asset (bet) has a well-defined termination point at which its value becomes certain. The absence of this property is one of the factors that has made it so difficult to test for rationality in the stock market. Since a stock is infinitely lived, its value today depends both on the present value of future cash flows *and* on the price someone will pay for the security tomorrow. Indeed, one can argue that wagering markets have a better chance of being efficient because the conditions (quick, repeated feedback) are those which usually facilitate learning. However, empirical research has uncovered several interesting anomalies. While there are numerous types of wagering markets, legal and otherwise, this column will concentrate on racetrack betting and lotto-type lottery games.[1]

Racetrack Betting Markets

The "market" at the racetrack convenes for about 20–30 minutes during which time participants place bets on any number of the six to twelve horses in the upcoming race. In a typical race, participants can bet on each horse, either to win, place or show (as well as "exotic" bets which depend on the combined outcomes of two or more horses). The horses that finish the race first, second or third are said to finish "in-the-money." All participants who have bet a horse to win realize a positive return on that bet only if the horse is first, while a place bet realizes a positive return if the horse is first or second, and a show bet realizes a positive return if the horse is first, second or third. There is a separate "pool" of money kept for each type of bet. Payoffs are determined in a "parimutuel" fashion, which means that the winning bets divide the money wagered on losing bets, less transactions costs.[2] The transactions costs consist of a fixed percentage t, which includes the "track take" and "breakage," the additional cost incurred because all returns per dollar bet are rounded down to the nearest five or ten cents. These transactions costs are substantial, typically in the range of 15–25 percent depending on the type of wager and the locale.

The proportion of the money in the win pool that is bet on any given horse can be interpreted as the subjective probability that this horse will win the race. By summing over many races, one can check what proportion of the horses with subjective probabilities between, say, .2 and .25 actually won races. The results of this analysis are impressive. Horses rated by the crowd as most likely to win (the "favorites") do win most often (about 1/3 of the time), and the correlation between subjective and objective probabilities is very high.[3] Apparently the bettors in these markets have considerable expertise.

[1] A future column may address other betting markets, such as NFL betting. Authors of recent papers on this topic are requested to send copies to Thaler.

[2] Since payoffs depend only on the final odds, bettors do not know potential payoffs when they bet. In Britain and some other places bookies accept bets on a fixed odds system where bettors are promised a certain payoff if their horse wins.

[3] See, for example, the studies by Weitzman (1965), Rosett (1965), Ali (1977), and Snyder (1978).

Does the high correlation between subjective and objective probabilities imply that the racetrack market is efficient? That depends on the definition of market efficiency. If we assume for the moment that all bettors are expected value[4] maximizers with rational expectations, then two definitions of market efficiency seem appropriate.

Market efficiency condition 1 (*weak*). No bets should have positive expected values.

Market efficiency condition 2 (*strong*). All bets should have expected values equal to $(1 - t)$ times the amount bet.

While the racetrack may be surprisingly efficient, there is substantial evidence that both of these conditions are violated. The most robust anomalous empirical regularity is called the *favorite-longshot bias*. Specifically, the expected returns per dollar bet increase monotonically with the probability of the horse winning. Favorites win more often than the subjective probabilities imply, and longshots less often. This means that favorites are much better bets than longshots. Indeed, extreme favorites, those with odds[5] of less than 3-10 (that is, with a greater than 70 percent chance to win) actually have positive expected values, in violation of condition 1.

Figure 1 (taken from Ziemba and Hausch, 1986) illustrates the favorite-longshot bias using data from most of the previously published studies (including over 50,000 races). Expected returns per dollar bet are plotted for horses at various market odds, using a transactions costs assumption of $t = 15.33$ percent, which applies in the state of California. The horizontal line indicates the point at which returns are the expected $.8467 = (1 - t)$. This occurs at odds of about 9-2 (that is, about a 15 percent probability of winning). For odds above 18-1 there is a steep drop in the expected return, with returns falling to only 13.7 cents per dollar wagered at 100-1. This implies that the typical 100-1 shot has real odds of about 730 to 1! Below 3-10 expected returns are positive, with returns of about 4-5 percent for the shortest odds horses. (This is partially explained by the $1.05 per dollar minimum payoff at nearly all U.S. tracks.) Although such overwhelming favorites are too rare to get very excited about, other profitable betting strategies are discussed below.

Another test of market efficiency is to compare the payoffs of equivalent bets. For example, most tracks offer a "daily double" bet which requires bettors to select the winners of the first two races. Suppose a bettor is considering buying a daily double ticket on horse A in the first race and B in the second race. Then an alternative betting strategy (called a parlay) would be to bet on A in the first race, and, if A wins, bet the proceeds on B to win the second race. Efficiency requires that the daily double payoff on A and B be the same as the parlay on A and B. This proposition has been tested by Ali (1979) and by Asch and Quandt (1987). The conclusion from these tests seems to be that daily double and parlay bets are priced reasonably efficiently relative to each other, though bettors should prefer the daily double because it offers lower transactions costs.

---

[4] Risk neutrality seems like a sensible initial assumption since most bettors probably wager a small portion of their total wealth. Other assumptions will be considered in the commentary section.

[5] Probabilities at the racetrack are traditionally quoted as "odds." If a horse has odds of "x to 1" then the implicit probability the horse will win is $(1 - t)/(x + 1)$.

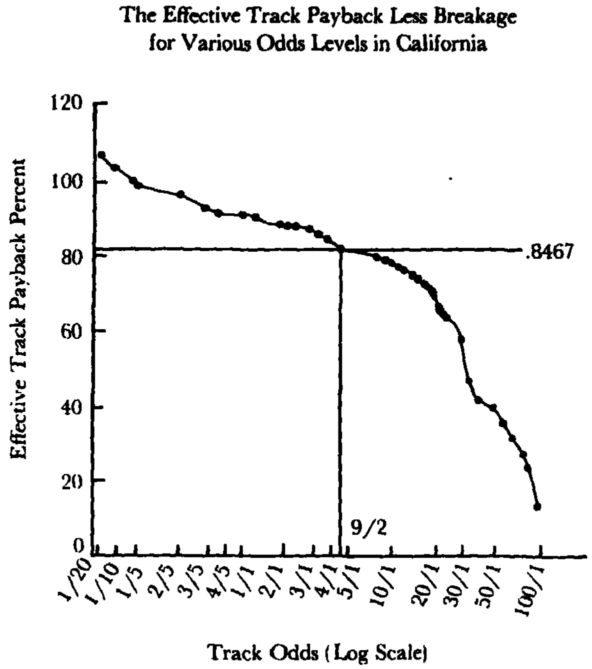


Fig. 1. The effective track payback less breakage for various odds levels in California

A similar test is possible using exacta betting in which a bettor must correctly pick the first and second place horses in the correct order. Just as the relative amounts bet on different horses in the win pool can be used to calculate implicit forecasts of the probability of winning, similar calculations for the exacta can be made using the so-called Harville (1973) formula. If q_i is the probability that horse i wins, then it is assumed that the probability that horse i is first and horse j is second is $q_i q_j/(1 - q_i)$. (Similarly, the probability that horse i is first, j is second, and k is third is $q_i q_j q_k/(1 - q_i)(1 - q_i - q_j)$.)[6] Asch and Quandt (1987) used the Harville formula to compare the subjective probability of winning implied by the betting in the win pool and the exacta pool. They found that the public did not bet in a mathematically consistent fashion. The implicit probabilities of winning for a given horse were often very different in the two pools.

Betting Strategies

Betting strategies at the track, as at the stock market, come in both fundamental and technical varieties. Fundamental strategies commonly are based on publicly available information used to "handicap" races. A bettor using a fundamental or handicapping strategy attempts to determine which horses, if any, have probabilities of winning (or placing or whatever) that exceed the market-determined odds by an amount sufficient to overcome the track take.[7] Technical systems require less information and use only current betting data. Bettors using a technical system attempt to find inefficiencies in the market and bet on such "overlays" when they have positive expected value. Most academic research has concentrated on the latter strategies.[8]

---

[6] The Harville formulas are quite accurate considering how little data they require. However, they tend to overestimate the probability that a low-odds horse will finish exactly second or third. More accurate estimation formulas are derived in Stern (1987), but these require data on all the horses in the race.

Hausch, Ziemba and Rubinstein (1981) (HZR) develop and test a strategy for betting in the place and show markets. They use the amounts bet in the win pool and the Harville formulas to calculate the probabilities of placing and showing for each horse based on the betting in the win pool. Using these methods they are able to identify horses that are underbet to place or show. The basic idea is to compare the proportion of the win pool bet on horse i with the amount of the place or show pool bet on horse i. If, for example, 40 percent of the bets in the win pool are on horse i, but only 15 percent of the bets in the place pool are bet on horse i, then it is profitable to bet on horse i to place. Such profitable betting opportunities typically occur 2 to 4 times per racing day. Empirical studies on two seasons of racing data indicate that significant returns on the order of 11 percent per bet are possible in the place and show markets.[9] This violates weak market efficiency. Moreover, publication of the system does not appear to have eliminated the profitable betting opportunities. At the recent 1987 Breeders' Cup races there were five "system" bets during the racing day. As luck would have it, all five bets paid off.

Ziemba and Hausch (1986) also developed similar techniques to identify and exploit inefficiencies in the exacta markets. The most frequent profitable bets have the favorite in second position. Their plots of the probability of winning and coming in second versus odds as shown in Figure 2, show that short odds horses have a substantial probability of coming in exactly second. The betting public might easily underestimate this chance. The other common profitable wagers are derived from the extreme favorite-longshot bias. Betting extreme favorites in the first position can thus yield profitable bets. The public wagers a considerable amount on these super horses, but not as much as they should. Combinations of longshots are almost never a good exacta wager; such bets typically return 10 to 30 cents on the dollar.

Asch, Malkiel and Quandt (1984, 1986) and Asch and Quandt (1986) investigated whether a drop in the odds late in the betting period might reflect inside information and thereby point to wagers that may have positive expected returns.

[7] A useful and insightful book on handicapping, based on a multitude of actual data, is Quirin (1979). The current state of the art in handicapping is described in Mitchell (1987) and Quinn (1987).

[8] Asch and Quandt (1986) did investigate whether the advice of professional handicappers ("touts") or computerized handicapping systems can be used to find profitable bets. They concluded that neither were very useful. See also Mitchell (1987).

[9] For details on the system see Ziemba and Hausch (1987). Hausch and Ziemba (1985) extend the HZR results to analyze the effects of differences in transactions costs, two-horse "entries" and multiple wagers, and provide accurate regression models for varying wealth levels and track "handles" (the size of the betting pool). They also investigate how many bettors have to follow the system for the market to become efficient.

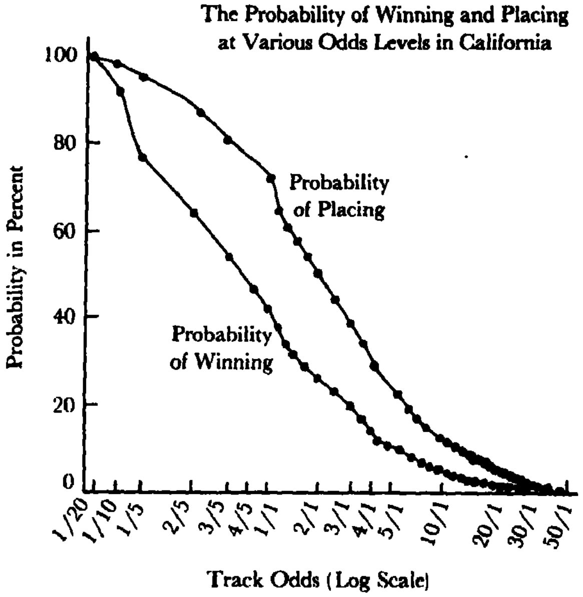


Fig. 2. The probability of winning and placing at various odds levels in California

Common racetrack folklore suggests that the smart money is bet late.[10] This is borne out by Asch, Malkiel and Quandt (1982) using data at various points in the betting cycle from 729 races at the Atlantic City Race Course. They found that for winning horses the final odds tend to be lower than the "morning line odds" (predicted odds by the track handicapper), whereas for horses finishing out of the money the final odds are much higher than the morning line odds. The later in the betting period, the more pronounced is the effect for the winners. The final odds for winners are 96 percent of the morning line odds, but for the money bet during the last eight minutes, the marginal odds are 82 percent of the morning line, and in the last five minutes they drop further to 79 percent. The final odds for losers are about 1.5 times the morning line odds. Asch and Quandt (1986) develop a logit model of the probability of winning, using the change in the odds during the last few minutes as one of the independent variables. The logit model is then used to search for profitable investment strategies. They could not find any profitable bets in the win pool, but they did find some in the place and show pools. Apparently, place and show betting on favorites whose odds have fallen in the last few minutes yields small profits. This is consistent with the Ziemba and Hausch (1987) results suggesting inefficiencies in the place and show pools.

[10] Asch and Quandt (1986) argue that this occurs because the smart money bettors want to withhold any information that their bets might convey. This implies that the public will bet more on a horse if the odds are *lower* than expected: an upward sloping demand curve! But if bettors have reservation odds for each horse, then they will also bet more on a horse if the odds are *higher* than expected. Probably the betting public contains a mixture of people using each strategy.

Cross Track Betting

A recent development in race track betting is the opportunity for bettors to wager at their home track on major thoroughbred races being run at another track. Cross track betting raises new and interesting questions about market efficiency. While arbitrage is made difficult by the high transactions costs and the absence of public telephones inside most race tracks, rational expectations would seem to imply that the odds at every track would be approximately the same. In fact, they frequently vary dramatically. For example, in the 1986 Kentucky Derby, the winner Ferdinand paid $16.80 for $2 at Hollywood Park in California where he had run often and was well known. He paid $37.40 at Aqueduct in New York, $79.60 at Woodbine in Toronto, $63.20 at Hialeah in Florida, and $90.00 at Evangeline in Louisiana.

While pure arbitrage may be difficult, profitable betting strategies are possible. Hausch and Ziemba (1987) have developed an optimal betting model for cross track betting under the assumption that final odds at all tracks are known in time to compute and place bets at each track. The essence of the system is to assume that the home track odds are accurate (after correcting for the favorite-longshot bias) and then to select a combination of bets at other tracks to exploit the inefficiencies. If the discrepancies in the odds at the various tracks are large enough (as they have been at some races), it is even possible to create a genuine arbitrage opportunity by betting on every horse at the track where the odds are best. Unfortunately, in the absence of a sophisticated communications system, these strategies are likely impractical (and possibly illegal). However, a Chicago commodities trader has developed and profitably used a workable one-track system using a portable television at the cross track. The bettor views the home odds when they are flashed on television, and then searches for overlays at the cross track.

Lotto Games

Lottery games date at least to biblical days. Israel was divided among the seven tribes by lot. Christ's robe was given to a lottery winner so it would not have to be cut. The Sistine chapel and its paintings were supported by lotteries. The Italian lottery has been running continuously since 1530. Lotteries are played in over 100 countries. Lotteries arrived in North America with the Pilgrims, and they were used to partially fund the new schools such as Harvard, Princeton and Yale. Later they were used to pay off debts of notables such as Thomas Jefferson. Extreme corruption led to their demise in the late 19th century, and they were banned in the United States and Canada. They resurfaced in 1964 in New Hampshire. In Canada they arrived to repay the debts from Expo 1967 in Montreal. Since then there has been explosive growth in popularity and sales. However, with an expected return of between 40 and 60 cents on the dollar, they are usually a poor investment for the rational investor.

Even with such low payout rates, it is possible to obtain positive expected value bets in lotto games. This occurs because not all numbers are equally popular with the public. The possibility of exploiting this pattern was first formalized by Chernoff

(1980) (and tested by some of his students) in the context of the Massachusetts numbers game. In this game the object is to pick a number from 0000 to 9999. If your number is drawn, then you share a portion of the total pool. A subsidiary prize is awarded if three numbers match. Chernoff found that certain numbers were unpopular: those with 0's, 9's, and to a lesser extent 8's. His theoretical analysis suggested that there were combinations with positive expected values, inducing some of his students to bet systematically on the "good" numbers. However, the students did not fare very well. First, over time the unpopular numbers became less advantageous, due to a combination of learning and simple regression. Second, they fell victim to the dreaded "gambler's ruin." The students' bankroll was not sufficient to wait out the time needed to have enough hits to generate substantial profits. Finally, they were unlucky: the unpopular numbers came up less often than would be expected.

The game that has attracted the most attention in North America is Lotto 6/49 or some similar variant. In this game one chooses six of forty-nine numbers and if they all match then one wins the jackpot. Lesser prizes are awarded for three to five matches. The probability of selecting the winning combination in this game is 1 in 13,983,816: if you play twice a week you can expect to win in 134,360 years, a long time horizon even for a rational economist!

Two features make the game interesting for the rational investor. First, as in the numbers game, some numbers are more popular than others. Second, if the grand prize is not won in a given drawing, it is carried over to the next week. Thus prizes can be enormous.[11] Ziemba and his co-workers (1986) have been studying whether these factors can produce favorable investment opportunities. Several estimation methods have been used to calculate the best numbers: simple counts of the frequency with which the numbers are picked, a regression of the log of payoffs on the winning numbers, and a sophisticated constrained maximum likelihood model. All lead to the same conclusion, namely that 15 to 20 of the numbers are quite unpopular. Moreover, the precise numbers are virtually the same from year to year. While there has been some learning over the years, so that these numbers are not quite as unpopular as they used to be, the unpopular numbers tend to remain unpopular. In fact, there are thousands of combinations of numbers that have expected returns over $1 *even when there is no carryover*. The expected value of betting the best numbers increases with the carryover and converges to about $2.25 per $1 for very large pools. The best numbers tend to be high numbers (non-birthdays) and those ending in 0's, 9's and 8's. According to the regression model, the twelve most unpopular numbers are 32, 29, 10, 30, 40, 39, 48, 12, 42, 41, 38, and 18 which tend to be 15 percent to 30 percent less popular than average. Using the marginal approach (those numbers chosen two standard deviations less than average), one finds the nineteen most unpopular numbers to be 40, 39, 20, 30, 41, 38, 42, 46, 29, 49, 48, 32, 10, 47, 1, 37, 28, 34 and 45. These numbers have edges from 26.7 percent down to 3.2 percent. The most popular number is 7, which is selected nearly 50 percent more often than the average number.

---

[11] In the U.S. lottery prizes are often announced using the accounting methods favored by university fund raisers and agents for professional athletes, namely undiscounted nominal dollars. The present value of the prize after taxes is typically about one-third of the announced value. In Canada, however, prizes are paid in cash and incur no Canadian tax liability.

The question remains—can you make money in the lotto games playing the unpopular numbers? While you can achieve an expected value of $2 per dollar bet, there may not be any arbitrage opportunity available. Consider a hypothetical carnival game with one million spokes. You pay $1 for a number between 1 and 1 million, and you get $2 million if your choice comes up. While you have an edge, the chance of winning is so small that you probably will go bankrupt before winning the jackpot. To analyze this problem one needs to have an adequate model of growth of wealth versus security of wealth. MacLean, Ziemba and Blazenko (1987) have developed such a model to investigate questions such as: Can a dynasty enhance its long-term wealth playing lotto games? The answer is that it can. With sufficiently small wagers it can increase its initial stake of, say, $10 million by tenfold before losing $5 million with probability arbitrarily close to one, but this process takes thousands of years even if they play all over the world. For one lotto game it will take them millions of years. A more interesting question for most of us is: Can a group or single investor use the unpopular numbers to become rich? This is even more difficult, especially if one wants low risk. It is easy to bankroll, as the optimal wager can be as low as 10 cents per week for one of ten syndicate members, but these aspiring millionaires are most likely to be residing in a cemetery when their distant heirs finally reach the goal. It is still best to play unpopular numbers—they have an edge and you will win 3 to 7 times the usual prizes should you hit—but you will expect to play a very long time before winning.

One of the most attractive aspects of lotto games is that the portion of the pool designated for the jackpot is carried over to the next draw when no one wins the jackpot. Indeed it is the prospect of winning a huge jackpot that is the main driving force behind the tremendous interest and sales of lotto tickets. Does it ever pay to buy all the numbers and hence "steal the pot"? Two conditions are necessary for this to be profitable. Roughly speaking these are: (1) a large carryover (in 6/49, $7.7 million); and (2) "not very many" tickets sold. While these conditions are unlikely to obtain, there are cases that can and have arisen in minor lotto games in Canada and elsewhere where it actually would have been a reasonable idea to buy the pot. It is important to stress, however, that even if the right conditions arose, buying the pot would entail enormous transactions costs since the tickets must be bought and redeemed one by one, and, you would have to hope that no one else tried to buy the pot at the same time (see Ziemba *et al.*, 1986, for details). Similar situations sometimes arise in exotic racetrack betting such as the "pick six" (pick the winners of six consecutive races) and related exotic bets. Substantial carryovers can exist in these pools, and making large wagers or actually buying the pot can be profitable. In fact, there are at least two major syndicates that actually try to do this, one of which made over $1 million last year.

Commentary

Racetrack betting

The racetrack betting market is surprisingly efficient. Market odds are remarkably good estimates of winning probabilities. This implies that racetrack bettors have considerable expertise, and that the markets should be taken seriously. Nevertheless, two robust anomalies are present: the favorite-longshot bias, and the inefficiencies of the place and show markets. How can these anomalies be explained?

Quandt (1986) has offered the following argument regarding the favorite-longshot bias (see also Rosett, 1965). The fact that bettors make wagers that are known to have negative expected value implies that they must be "locally" risk seeking. This implies that the usual risk-return relationship will be reversed. In equilibrium, investments (bets) with high variance will have lower average returns than investments with low variance. While this argument is logically consistent, we feel that it is not a satisfactory explanation of the observed behavior. The crucial issue is whether the inference that bettors are risk-seeking is a reasonable one to draw from the fact that they are at the track betting.

What does it mean to be "locally risk seeking"? Recognizing that most racetrack fans, including themselves, purchase insurance, Asch and Quandt (1986) suggest that the utility of wealth function may have the shape proposed by Friedman and Savage (1948), namely concave below the current wealth level and convex above it. While this assumption can explain why racetrack bettors also purchase insurance, it is surely not an adequate explanation for bettors' other behavior such as investing. We venture a guess that when it comes to retirement saving, Professors Asch and Quandt would not be willing to accept a lower mean return in order to obtain a higher level of risk. Indeed, having read their coauthor's book on the stock market (Malkiel, 1985) we guess that when it comes to investing, many racetrack bettors display concave utility functions. Thus the term "locally risk seeking" may apply to racetrack bettors, but only if the term "locally" refers to physical location rather than wealth level![12]

It is true that racetrack fans go to the track to bet—watching a horse race is just not that much fun if you do not have a rooting interest. The real question is to what extent we can explain racetrack betting with the assumptions of rational expectations, expected utility maximization, and a convex utility of wealth function. Consider some stylized facts about racetrack bettors: First, most bring a stake that represents a small portion of their wealth. The average amount bet per person in 1985 was about $150 for the day. The median is surely lower. Second, they allocate that stake over the course of the betting day, intending to bet on nearly every race, unless they run out of money before the day ends. Third, groups of friends that attend the track together rarely bet among themselves, although they could thereby guarantee a zero sum game for the group and increase variance as much as they wanted. Are these facts consistent with maximization of a convex utility of wealth function?

---

[12] The more basic question is whether individuals display a consistent "trait" that can be captured in an index of risk aversion or risk seeking. Psychologists have found that most such traits are highly context specific, and risk taking is no exception. As Paul Slovic (1972, p. 795) has commented: "Although knowledge of the dynamics of risk taking is still limited, there is one important aspect that has been fairly well researched–that dealing with the stability of a person's characteristic risk-taking preferences as he moves from situation to situation. Typically, a subject is tested in a variety of risk-taking tasks involving problem solving, athletic, social, vocational, and pure gambling situations. The results of close to a dozen such studies indicate little correlation, from one setting to another, in a person's preferred level of risk taking."

Another fact that is difficult to explain within this framework is the tendency (first pointed out by McGlothlin, 1956) for the favorite-longshot bias to become more pronounced for the last couple of races of the day. Most observers (McGlothlin, 1956; Kahneman and Tversky, 1979; Asch and Quandt, 1986) seem to agree on the cause. Bettors on average are losing toward the end of the day. They would like to go home a winner, but do not want to risk losing much more money. Therefore, they bet on longshots in an attempt to break even for the day. Notice that this behavior is hard to explain within a Friedman-Savage framework. Why should a reduction in wealth increase the tendency for risk seeking?

We feel that a more promising way of modeling race track betting (and other gambling behavior) is to introduce the concept of *mental accounting* (Kahneman and Tversky, 1984; Thaler, 1985). The key assumption of mental accounting is that people adopt mental accounts and act as if the money in these accounts is not fungible. To get the feel of mental accounting, consider another thought experiment. A set of identical twins Art and Bart (with identical wealth levels) is at the race track, contemplating their bets for the last race of the day. Art has lost $100 betting so far, though he has another $100 in cash with him. Bart is even in the betting so far, but between races he read the financial page in the newspaper and discovered that a stock in which he holds 100 shares went down one point the previous day.

Notice that both twins have lost $100, and thus any wealth-based explanation of their betting behavior must predict that they will make similar bets. However, in a mental accounting formulation, Art is behind in the race track account while Bart is even; thus they might well bet differently. (See Thaler and Johnson (1986) for evidence consistent with this view.) Once the concept of mental accounting is introduced, then it becomes much easier to understand how an individual can be risk neutral or risk seeking at the racetrack but risk averse with respect to retirement savings.

As for the favorite-longshot bias, many behavioral factors are probably at work for different reasons. (1) Bettors might overestimate the chances that the long shots will win. (2) As in Kahneman and Tversky's (1979) prospect theory, bettors might overweight the small probability of winning in calculating the utility of the bet. (3) Bettors may derive utility simply from holding a ticket on a longshot. After all, $2 is a cheap thrill. (4) It is more fun to pick a long shot to win than a favorite. It is hard to claim much credit for predicting that a 1-5 favorite will win (much less place or show), but if a 20-1 longshot comes through, considerable bragging rights will have been earned. (5) Some bettors may choose horses for essentially irrational reasons, like the horse's name. Since there is no possibility of short sales, such bettors can drive the odds down on the worst horses, with the "smart money" simply taking the better bets on the favorites.

The fact that the place and show pools seem to be less efficient than the win pool is also an interesting observation. One important factor may simply be that these bets are more complicated. For example, the payoff to a bet to show depends not only on the chance the horse will be in the money, but also on which other horses are in the money and how much has been bet on each. (The greater is the share of the money that has been bet on the horses finishing in the top three positions, the smaller is the payoff.) Bettors might prefer simple bets to complicated bets,[13] or they might simply have difficulty determining when an attractive bet occurs in place and show pools.

One important conclusion we draw from this analysis is that modeling gambling behavior is complicated. Bettors' behavior seems to depend on numerous factors such as how they have done in earlier races, and which bets will yield the best stories after the fact. We should emphasize that these complications apply with equal force to investment behavior. As Merton Miller has said (1986, S467), "[to many individual investors] stocks are usually more than just the abstract 'bundle of returns' of our economic model. Behind each holding may be a story of family business, family quarrels, legacies received, divorce settlements, and a host of other considerations almost totally irrelevant to our theories of portfolio selection. That we abstract from these stories in building our models is not because the stories are uninteresting but because they may be too interesting and thereby distract us from the pervasive market forces that should be our principal concern." While we sympathize with Miller's self-control problem—we also find the stories irresistibly interesting—we feel that to understand the market forces one must enrich the models to incorporate more than the "bundle of returns." Even professional portfolio managers seem more concerned with beating the S&P index than with maximizing returns. In fact, we suspect that portfolio managers trailing the market in the 4th quarter may behave much like the racetrack bettors who bet on longshots when behind at the end of the day.

Lotteries

What can economic theory say about lotteries? Given the dreadful payout rates, one prediction might be that no one will purchase lottery tickets. However, it is easy to rationalize the purchase of a lottery ticket by saying that for a dollar purchase, the customer is paying 50 cents for a fantasy. That's a pretty good deal. The existence of popular and unpopular numbers is more difficult to rationalize. It seems that economic theory yields the following paradoxical prediction: No one will choose the most popular numbers.

To understand this phenomenon it is useful to point out that lotteries in North America did not become popular until New Jersey introduced a game which allowed players to choose their own numbers. The popularity of this feature seems to be explained by what psychologist Ellen Langer (1975) has called "the illusion of control." Even in purely chance games, players feel they have a better chance to win if they can control their own fate, rather than have it determined by purely "chance" factors. For example, Langer found that subjects in her experiments were more reluctant (charged a higher price) to give up a lottery ticket they had selected themselves, than one selected at random for them.

---

[13] For a similar example in the finance literature, see Elton, Gruber, and Rentzler (1982).

A news story provides a vivid example of the illusion of control (and the confusion of skill and chance). One year, the winner of the Christmas drawing for the Spanish National Lottery, the "El Gordo," was interviewed on television. He was asked: "How did you do it? How did you know which ticket to buy?" Our winner replied that he had searched for a vendor who could sell him a ticket ending in 48. "Why 48?" he was asked. "Well, I dreamed of the number seven for seven nights in a row, and since seven times seven is 48..."[14]

■ *The authors wish to thank, without implicating, Peter Asch, Werner De Bondt, Donald Hausch, Richard Rosett, and Richard Quandt for helpful comments on an earlier draft. Financial support was provided by the Alfred P. Sloan Foundation.*

References

Ali, Mukhtar, M., "Probability and Utility Estimates for Racetrack Bettors,"*Journal of Political Economy*, 1977, *85*, 803–15.

Ali, Mukhtar, M., "Some Evidence of the Efficiency of a Speculative Market," *Econometrica*, 1979, *47*, 387–92.

Asch, Peter, and Richard E Quandt, "Efficiency and Profitability in Exotic Bets," *Economica*, August 1987, *59*, 278–98.

Asch, Peter, and Richard E. Quandt, *Racetrack Betting: The Professors' Guide to Strategies.* Dover, MA: Auburn House, 1986.

Asch, Peter, Burton G Malkiel, and Richard E. Quandt, "Market Efficiency in Racetrack Betting," *Journal of Business*, 1984, *57*, 65–75.

Asch, Peter, Burton G. Malkiel, and Richard E. Quandt, "Market Efficiency in Racetrack Betting: Further Evidence and a Correction," *Journal of Business*, 1986, *59*, 157–60.

Asch, Peter, Burton G. Malkiel, and Richard E. Quandt, "Racetrack Betting and Informed Behavior," *Journal of Financial Economics*, 1982, *10*, 187–94.

Chernoff, Herman, "An Analysis of the Massachusetts Numbers Game," Cambridge, Massachusetts, MIT Department of Mathematics, Technical Report No. 23, November 1980.

Elton, E., M. Gruber, and J. Rentzler, "Intra Day Tests of the Efficiency of the Treasury Bills Futures Market," Working paper No. CSFM-38, Columbia University Business School, October 1982.

Friedman, Milton, and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, August 1948, *56*, 279–304.

Harville, David A., "Assigning Probabilities to the Outcome of Multi-Entry Competitions," *Journal of the American Statistical Association*, 1973, *68*, 312–16.

Hausch, Donald B., and William T. Ziemba, "Cross Track Betting on Major Stakes Races," Vancouver: University of British Columbia, Faculty of Commerce Working Paper No. 975, June 1987.

Hausch, Donald B., and William T. Ziemba, "Transactions Costs, Extent of Inefficiencies, Entries and Multiple Wagers in a Racetrack Betting Model," *Management Science*, 1985, *31*, 381–94.

Hausch, Donald B., William T. Ziemba, and Mark Rubinstein, "Efficiency of the Market for Racetrack Betting," *Management Science*, 1981, *27*, 1435–52.

Kahneman, Daniel, and Amos Tversky, "Choices, Values, and Frames," *American Psycholo-*

---

[14] This example is cited in a forthcoming book on decision making by Jay Russo and Paul Schoemaker. We are grateful they shared it with us.

gist, 1984, *39*, 341–50.

Kahneman, Daniel, and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 1979, *47*, 263–91.

Langer, Ellen J., "The Illusion of Control," *Journal of Personality and Social Psychology*, 1975, *32*, 311–328.

McGlothlin, William H., "Stability of Choices Among Uncertain Alternatives," *American Journal of Psychology*, 1956, *69*, 604–15.

MacLean, Leonard, William T. Ziemba, and George Blazenko, "Growth versus Security in Dynamic Investment Analysis," University of British Columbia, Faculty of Commerce and Business Administration, mimeo, 1987.

Malkiel, Burton G., *A Random Walk Down Wall Street*. New York: Norton, 1985.

Miller, Merton H., "Behavioral Rationality in Finance: The Case of Dividends," *Journal of Business*, October 1986, *59*, S451–68.

Mitchell, Dick, *A Winning Thoroughbred Strategy*. Los Angeles: Cynthia Publishing Co., 1987.

Quandt, Richard E., "Betting and Equilibrium," *Quarterly Journal of Economics*, 1986, *101*, 201–207.

Quinn, James, *The Best of Thoroughbred Handicapping 1965–1986* New York: Morrow, 1987.

Quirin, William L., *Winning at the Races: Computer Discoveries in Thoroughbred Handicapping*. New York: Morrow, 1979.

Rosett, Richard N., "Gambling and Rationality," *Journal of Political Economy*, 1965, *73*, 595–607.

Slovic, Paul, "Psychological Study of Human Judgment: Implications for Investment Decision Making," *Journal of Finance*, 1972, *27*, 779–99.

Synder, Wayne W., "Horse Racing: Testing the Efficient Markets Model," *Journal of Finance*, 1978, *33*, 1109–18.

Stern, Hal, "Gamma Processes, Paired Comparisons and Ranking," Ph.D. dissertation, Department of Statistics, Stanford University, August 1987.

Thaler, Richard H., "Mental Accounting and Consumer Choice," *Marketing Science*, Summer 1985, *4*, 199–214.

Thaler, Richard H., and Eric Johnson, "Hedonic Framing and the Break-Even Effect," unpublished manuscript, Cornell University, Johnson Graduate School of Management, 1986.

Weitzman, Martin, "Utility Analysis and Group Behavior: An Empirical Study," *Journal of Political Economy*, 1965, *73*, 18–26.

Ziemba, William T., and Donald B. Hausch, *Betting at the Racetrack*. Vancouver and Los Angeles: Dr. Z. Investments, Inc., 1986.

Ziemba, William T., and Donald B. Hausch, *Dr. Z's Beat the Racetrack*. New York: William Morrow, 1987.

Ziemba, William T., Shelby L. Brumelle, Antoine Gautier, and Sandra L. Schwartz, *Dr. Z's 6/49 Lotto Guidebook*. Vancouver and Los Angeles: Dr. Z. Investments, Inc., 1986.

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 175–184

Correspondence

To be considered for publication in the Correspondence section, letters should be relatively short, generally less than 1000 words, and should be sent to the journal offices at the address appearing inside the front cover. The editors will choose which letters will be published, and all published letters will be subject to editing for style and length.

On "Symposium on Tax Reform," Summer 1987, pp. 7–119.

While the symposium on the 1986 Tax Act was highly informative, one aspect of the discussion could stand some comment: that dealing with the basis of comments about tax "equity," or justice. Three papers in the symposium deal directly with justice issues, and some of the others deal tangentially with them. In these, all of the critical comments dealing with taxation "equity" are based upon an ethical criterion which, while important in the thinking of contemporary public finance scholars, is hardly the only criterion that could be chosen. Other ethical yardsticks could be employed which arguably are just as reasonable, and when this is done, judgments about the justice of many of the tax provisions in the U.S. tax system can be different from the ones made by the writers in the symposium.

The ethical tradition I refer to is one in which income comprehensively defined (the value of all consumption plus net change in asset values), considered to be the best proxy for overall economic power, becomes the main or only basis of equity comparisons. This standard is especially evident in the articles by Joseph Pechman, by Richard Musgrave, and by Charles McLure and George Zodrow. Differences in source or use of income are to be severely downplayed, if not neglected altogether. But obviously, any number of ethical considerations are not very well represented by income by itself, no matter how broadly defined. Some obvious examples of alternative goals include reward for effort and/or ability, maintaining a viable federal system, protecting individual liberties, fostering decent life changes for all individuals, avoiding excessive concentrations of wealth, allowing women with children the opportunity to work, and structuring tax payments according to the perceived benefits persons receive from the existence of government.

What difference might it make in how we pass judgment on the 1986 Act if we take this broader view? A few examples come readily to mind:

Charging higher tax rates on income from inherited wealth follows from the social goal that persons ought to be rewarded for their *contributions*, and that this is the source of the economic incentives which are truly meaningful. The same is true for higher transfer taxes on inherited wealth.

Private charitable contributions are an extremely valuable method of achieving public good objectives without government intervention. However, the lowering of top marginal rates in the tax reform act increased the tax price to high-income persons of making charitable contributions by 44 percent. This, plus disallowing the special deduction for persons using the standard deduction and taxing the capital gain portion of appreciated assets at 21 percent, will have a substantial adverse affect on such contributions.

The Tax Reform Act did little to help women achieve equal economic opportunity in American society, especially women who are single parents. A generous credit for day-care expenses is a "leakage" which could in one stroke do much to overcome the economic disadvantages faced by women. The Act eliminated some credits for child-care expenditures which had existed previously in the tax law.

Despite its better treatment of poor people generally, the tax reform act left many hard-working persons who are near-poor worse off, mainly because of the elimination of the extra deduction for working mothers and of tax credits for day-care expenses. Many families with children where both parents work having incomes in the $20,000 to $30,000 range will have greater tax liabilities under the new bill.

It is a mistake not to consider taxation with respect to *all* important social goals, not just those that correlate highly with income comprehensively defined (and many do not). Tax students before the 1930s would have found nothing untoward in this view, and there is no reason why their *ethics* are inferior to positions taken in the late 20th century. To be sure, fine-tuning the tax code may introduce undue complexity, but the benefits of provisions need to be considered together with the complexity losses directly, not according to some academic ideal standard.

Herbert J. Kiesling
Professor of Economics and Public-Environmental Affairs
Indiana University
Bloomington, Indiana

On "Evaluating Uncertain Evidence with Sir Thomas Bayes: A Note for Teachers," by Steven C. Salop, Summer 1987, pp. 155–160

Since I am a law teacher and as Professor Salop's appointment is in a law school, I am especially troubled by the thought that his note is aimed at law teachers, for we are, for the most part, a gullible lot, easily bamboozled by the assertions of economists, such as Salop's debatable claims: "How much weight to place on imperfect evidence is a problem of statistical interference [*sic*]. The correct statistical methodology of combining such evidence is ... Bayes Theorem ... " In the legal literature, at least, the debate centers around the truth of these two claims.

Law teachers will undoubtedly forgive Professor Salop for a certain lack of realism in stating his hypothetical case—in reality, the Blue Cab company's insurance carrier would settle the case and then Mr. Doe would successfully sue his former attorneys for malpractice for not having asked for damages for pain and suffering and punitive damages and the Blue Cab company would discharge its attorneys for stipulating that that "company would be liable for injuries suffered in such an accident, if the owner of the offending taxi can be identified," for that stipulation amounts to a confession of liability if the Green Cab company (or anyone else) is identified as the owner.

What we should not forgive is Salop's failure to make explicit his assumption that ignorance is the equivalent of evidence; for the fact that there are 85 green cabs and 15 blue cabs in an urn does not justify the conclusion that when one draws out a cab the odds are 85 to 15 that that cab is going to be green. That conclusion is only acceptable if one assumes, as Salop must, that the cabs are randomly distributed within the urn.

The logs of the cab drivers on the night of March 1, the police report of the accident, the past accident rates of the two cab companies, the physical condition of the cabs on March 2, and much, much more would be needed to get a decent idea of the distribution of blue and green cabs on the night of the accident. Without such information, the number of cabs registered to the two cab companies is without probative value; with such information, the number of registered cabs per company is likely to be irrelevant.

What Professor Salop, like many economists (and many lawyers), fails to recognize is that the purpose of a trail is to reach a just result by just means, not to determine truth, and certainly not to determine probabilities by statistical means. The concept of procedural justice—of due process—is not one that can be captured by a statistical model. If Professor Salop's case were to be tried, rather than settled, it would be a jury, not a judge, that would determine whether the Blue Cab company is liable, though, as Salop decries, juries are even less likely than judges to be willing to substitute the mechanical calculus of Bayes Theorem for their own judgment. But the right to a jury trial is enshrined in our federal and state constitutions, which strongly suggests that deciding the weight of imperfect evidence—and evidence is never perfect —is much more than a problem of statistical inference.

If Professor Salop is addressing teachers of economics rather than of law, the defects in his note are in some ways more disturbing. Economics teachers surely will not be taken in by his implicit assignment of subjective probabilities to issues about which we have no knowledge, especially as he recognizes in another context that the number of negligent cabs owned by a cab company is not a monotonic function of the number of cabs owned by the company. But economics teachers may not be aware that Professor Salop's description of the legal system is a grotesque parody of the way it actually works. I admit that some legal academics, and one or two recently appointed federal judges, might accept this parody as fact, but that is only an example of the social cost of the law and economics movement with its irrational tendency to apply, without the slightest empirical justification, abstract and trivial economic models to complex systems and, worse, to issues of justice and of social values.

Peter D. Junger
Professor of Law
Case Western Reserve University
Cleveland, Ohio

Response from Steven C. Salop: Professor Junger would not allow the Registrar's testimony into evidence because it is too imprecise. He would want to know the logs of the cab drivers, the police report of the accident, the past accident records of the taxi firms and the physical conditions of the taxis.

I agree with Professor Junger that more evidence is better than less and I would expect the parties to introduce the additional information he suggests. Bayesian methods are useful regardless of the particular information. My point was simply that these various pieces of information should be balanced by the fact finder—either judge or jury—using Bayesian methods and intuitions, not simply by what feels right and just.

Further, I must say that I was somewhat surprised to learn that my note qualified as one more example of the "social cost of the law and economics movement with its irrational tendency to apply, without the slightest empirical justification, abstract and trivial economic models to complex systems and, worse, to issues of justice and of social values." My goals were far less grandiose than that. I merely wanted to teach the logic and psychological limitations of Bayes Law. Perhaps Professor Junger has me mixed up with someone else. As to the easy gullibility of law teachers, I agree, of course. Indeed, Professor Junger stands as one probative, though surely imperfect, source of evidence on this point.

Finally, I wanted to fill in some citations that did not appear in my original note. While much of the literature was cited in great detail in the companion article by Mark Machina, the key articles should have been noted in my paper, too. This oversight was unfortunate because both the taxi cab problem and the Asian disease problem, formulated originally by Daniel Kahneman and Amos Tversky, are classics. Students who read about these problems should know who thought them up originally.

The correct citations are as follows. The Asian disease problem is analyzed and experimental results are discussed in Tversky and Kahneman, "The Framing of Decisions and the Rationality of Choice," *Science* (1981), pp. 453–58, as well as in the articles mentioned by Machina in his paper. The taxi cab problem is discussed in detail in Kahneman, Slovic and Tversky, *Judgement Under Uncertainty: Heuristics and Biases* (1982), pp. 153–160. Both problems were developed somewhat earlier.

Without these two brilliant problems, my teaching note would have had nothing to teach. Of course, it is interesting that, once one begins to investigate the pedigree of classic ideas, properly allocating credit becomes more complicated. Scholars often talk about an idea being "in the air." These decision problems are no exception. For example, Lawrence Tribe's 1971 article cited in my note poses a hypothetical case in which the plaintiff is run down by a blue bus and can prove only that defendant operates four-fifths of all the blue buses in town. Better yet, as Tribe points out, something close to this hypothetical actually occurred in a the case of *Smith v. Rapid Transit, Inc.*, 58 N.E. 2d 754 (1945).

Nonetheless, it is quite clear that Kahneman and Tversky deserve the credit for developing this research area and, thus, the honor of attribution. Without the research they started, economists still might be assuming that consumers are more like rats than humans.

The ultimate compliment to a scholar is having an idea named for its creator. Tobin's q, Occam's razor and Viner's draftsman demonstrate the range of this honor. Unfortunately, Kahneman–Tverskyism does not have an easy ring. Accordingly, I suggest, first, that the authors of *Principles of Economics* texts start including this research and, second, that economists refer to this type of human behavior as "KT Judgement Bias."

On "Choice Under Uncertainty: Problems Solved and Unsolved," by Mark J. Machina, Summer 1987, pp. 121–154.

Mark Machina's article on the uncertainty literature provided readers like myself whose research interests lie elsewhere a much better understanding of the current issues and debates in research on uncertainty. This comment sets forth two issues suggested by the Machina article, but not directly confronted by his discussion.

The first issue concerns a particular feature of what Machina calls "non-expected utility models of preferences." Machina lists five of the available generalized forms (equations 4 through 8), and points out (p. 133) that "each requires a different set of conditions on its component functions" for the form to display a number of desirable properties such as "stochastic dominance preference, risk aversion/risk preference, comparative risk aversion, etc."

A question unanswered by Machina's discussion of these general forms has to do with the ability or failure of each form to resolve what to some of us is a major disadvantage of standard expected utility theory: its inability to provide an intuitively appealing explanation for the existence of individuals who both gamble and buy insurance. Machina cites Friedman and Savage's well-known 1948 article as showing "how a utility function which was concave at low wealth levels and convex at high wealth levels could explain the behavior of individuals who both incur risk by purchasing lottery tickets as well as avoid risk by purchasing insurance." But this particular "solution" to the problem of the existence of individuals who both gamble and buy insurance has always seemed inadequate to a sizable proportion of the profession, since it requires a very specialized and not particularly intuitively appealing assumption (the terms "convoluted" and "Ptolemaic epicycles" spring to mind) about the shape of the typical individual's utility function.

Surely a convincing theory should be consistent with this major fact about real world behavior. If any of the new forms were "automatically" consistent with such behavior, that would seem to be a big point in their favor. Lack of such consistency would count heavily against them. My question for Machina is whether any or all of the more general utility function forms he identifies are "automatically" consistent with the behavior of an individual that both gambles and buys insurance; the term "automatically" is meant to exclude the incorporation of additional specialized nonintuitive restrictions (analogous to Friedman–Savage type "convolutions") imposed solely to make the utility function consistent with the joint existence of gambling and insurance.

A second issue concerns the way in which expected utility theory has been confronted with apparently anomalous experimental evidence. A well-known and seemingly widely-accepted methodological dictum in economics, associated with Milton Friedman's famous methodology essay, is that theories are appropriately tested by confronting their predictions, not their *assumptions*, with empirical evidence. As I understand Machina's discussion, the experiments that cast doubt on the descriptive correctness of the independence axiom as an assumption also cast doubt on the predictions of expected utility theory, and I infer that Machina is telling us that this predictive failure is the basis for criticizing the standard expected utility framework. This raises a number of interesting questions:

(1) Suppose my inference is wrong, and the theory *actually is* being rejected based on a rejected assumption, rather than failed expected utility predictions. If the answer is true, isn't this a methodological error, and why has such an error persisted?

(2) Suppose my inference is correct, and the theory is being rejected because of failed predictions. A striking feature of this situation is the apparent ability of the literature to attribute the predictive failure back to a *specific* assumption. It is not obvious why demonstrating that one of several assumptions is "wildly inaccurate" (to use Friedman's words) is a demonstration that this *particular* wild inaccuracy is responsible for predictive failure. Perhaps preserving this wildly inaccurate assumption and changing another (perhaps untested) assumption would produce a theory that predicted satisfactorily. If the independence axiom is in fact an assumption, and one of several, how can it be "equivalent" to a derived *result* of the theory: the proposition that expected utility calculations are both appropriate and linear in the probabilities? Is there something special about the structure of expected utility theory that allows us to attribute to a particular axiom (assumption) the *sole* blame for the theory's predictive failure?

(3) There are also puzzles associated with the implications for theory construction of directly testing axioms. Consider a theory which had n axiomatic assumptions. A subset k of these axioms was testable by laboratory experiments, while $n - k$ were not. The theory's record at predicting was undistinguished, but better than available alternatives. Laboratory experiments show that c of the testable axioms are refuted,

while $k - c$ are not. Does this laboratory information about the "validity" of the axioms *by itself* provide particular guidance to the theorist about how he should modify the theory to try to improve its predictions? It is not obvious to me why it would provide such guidance (based partly but not entirely on the Friedmanesque view about wildly inaccurate assumptions discussed above), but I wonder what Machina's reaction would be about whether this laboratory information provides guidance?

Robert S. Goldfarb
Professor of Economics
George Washington University
Washington, D.C.

Response from Mark J. Machina. Professor Goldfarb has raised two issues in connection with my article "Choice Under Uncertainty: Problems Solved and Unsolved." The first is whether the non-expected utility models described there can correct what he sees as expected utility's "inability to provide an intuitively appealing explanation for the existence of individuals who both gamble and buy insurance." The answer to this question depends on what we mean by "gamble and buy insurance." If we mean the simultaneous purchase of actuarially unfair insurance and actuarially unfair lottery tickets at a *given* wealth level x_0, then the problem with the expected utility model is that it is too *weak* to imply (i.e. to necessitate) such behavior without additional assumptions on the shape of the von Neumann-Morgenstern utility function $U(\cdot)$ over final wealth levels, namely that $U(\cdot)$ be concave for wealth levels below x_0 and convex for wealth levels above it, *a la* Friedman and Savage.[1] Since the non-expected utility models I discussed are designed to accommodate empirical violations of the expected utility model and are thus strictly *weaker* than it, they too require additional assumptions in order to actually necessitate such behavior,[2] and hence do not differ from expected utility in this respect.

However, the more interesting notion of simultaneous gambling and insurance is the observed propensity of individuals to purchase both unfair insurance and unfair lottery tickets at *all* wealth levels.[3] In this case the problem with the expected utility model is that it is too *strong*, in the sense that it rules out such behavior. After all, it is impossible for a stable (i.e. fixed) function $U(\cdot)$ over final wealth levels to be concave below x and convex above x for *every* wealth level x. The most well known attempt to weaken the model to accommodate this observation is that of Markowitz (1952), who retained the expected utility property of "linearity in the probabilities" but dropped the assumption of a stable utility function over final wealth levels. Markowitz hypothesized that at any given moment there would be some von

---

[1] Milton Friedman and Leonard J. Savage, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, August 1948, *56*, 279-304.
[2] See for example Mark J. Machina, "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 277-323, in particular Section 4.2.
[3] See Friedman and Savage (1948, Section 2) and Harry Markowitz, "The Utility of Wealth," *Journal of Political Economy*, April 1952, *60*, 151-158.

Neumann–Morgenstern utility function over final wealth whose expectation represented preferences over probability distributions, but that changes in the individual's current wealth would cause this function to shift horizontally, so that (for example) the same individual may have different attitudes regarding a 50:50 chance of ending up with a final wealth of $100,000 or $300,000 versus a sure chance of ending up with $200,000, depending upon whether this choice is viewed from a current wealth position of $50,000 or a current wealth position of $500,000.

The newer non-expected utility models provide us with an alternative (and I feel superior) approach to this problem, namely retaining the stability of preferences but dropping linearity in the probabilities.[4] In Machina (1982, Section 4.5) I presented an example of a fixed non-expected utility preference function over final wealth distributions which would purchase unfair insurance at any initial wealth level (provided it is not *too* unfair) and also purchase unfair (though not too unfair) lottery tickets at any initial wealth level, provided the prize is high enough. This preference function also avoids symmetric fair gambles around any wealth level, assigns a finite certainty equivalent to the St. Petersburg (or any generalized St. Petersburg) gamble at any wealth level, and has indifference curves in the triangle diagram which always "fan out." Although this example cannot accommodate *all* of the empirical violations of the expected utility model[5] (no current model can), it presents a strong case for dropping linearity in the probabilities instead of the more fundamental assumption of stable preferences.

Goldfarb's second issue concerns the manner in which the expected utility model has been tested. He asks whether these empirical violations of transitivity, the independence axiom, etc. aren't in fact tests of the *assumptions* of the expected utility model rather than tests of its *predictions*; and if so, isn't this a methodological error, at least according to the methodology proposed by Friedman?[6] Although transitivity and the independence axiom are indeed necessary for the derivation of the expected utility model, they are also logical *implications* of the expected utility hypothesis, and hence fair game to test as predictions.

Finally, Goldfarb asks how we can attribute specific empirical violations to particular assumptions of the expected utility model. In the case of departures from linearity in the probabilities the guilt clearly lies with the independence axiom, since most non-expected utility models which satisfy all of the expected utility assumptions *except* this axiom can satisfactorily accommodate these observations. On the other hand, as seen in my section "The Preference Reversal Phenomenon," researchers sometimes *do* differ on exactly which (explicit or implicit) assumption is being violated by a given class of observations with very different implications for the types of alternative models they construct. When this happens, it's the job of theorists and experimentalists to be more explicit about both the assumptions and the predictions of their models. The most systematic and careful examinations of the theoretical implications and/or empirical validity of the component assumptions of the expected utility model are unquestionably those of Fishburn (theoretical) and MacCrimmon and Larsson (empirical).[7]

---

[4] After all, the evidence suggests that preferences are not linear in the probabilities even at a *given* initial wealth.

[5] Since it is transitive it cannot explain the preference reversal phenomenon, and since it is defined over ultimate wealth distributions it cannot explain framing effects.

[6] Milton Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

I would like to thank the Center for Advanced Study in the Behavioral Sciences for financial support, and to apologize for an unfortunate printing error in the original article: the five lines of text on page 131 *should appear after line* 18 *on the following page.*

On "Anomalies," by Richard Thaler, Summer 1987, pp. 197–201

The editors of the *Journal of Economic Perspectives* are to be congratulated for their decision to publish a regular column on anomalies in economic science. To this insight was added inspiration in the decision to ask Richard Thaler to write and edit the column. It promises to be a delightfully lively column under Thaler's pen. I do, however, wish to make a general point about the concept of anomaly.

Anomaly is not synonymous with disconfirmation. Anomaly is the *simultaneous* presence of disconfirming *and* confirming evidence. Furthermore, for the scientist to know what is to be done about the anomaly it is important to be well versed in all the evidence, both confirming and disconfirming. This is because the scientific task is either to (1) resolve the anomaly in terms of an appropriate reevaluation of the implementational subhypotheses (initial conditions) of the existing theory or, failing this, (2) extend or generalize the theory so that its increased empirical content includes the anomaly *as well as* the previous confirming evidence, and also has *new testable implications* not derivable from the old theory.

Case (1) was illustrated by the discovery of perturbations in the perihelion of Uranus. This observation could be interpreted as falsifying evidence for Newtonian physics. But of course all empirical tests of a theory depend upon often unarticulated subsidiary supporting hypotheses (the *ceteris paribus* clause). In this case the supporting hypothesis (among others) was that there was no unknown new planet in the system that accounted for the deviant motion of Uranus. By postulating such a planet, Newtonian scientists were able to account for the deviations, and subsequently observational astronomers discovered the culprit, Neptune. Case (2) is illustrated by the fact that Mercury showed a similarly deviant orbit that could not be accounted for by Newtonian mechanics. So, Newtonian scientists, flush with the Neptune success, computed the properties of a "10th planet" (or planets) nearer the Sun, that could account for the observations of Mercury's path, and boldly named "it" Vulcan. But alas, no one has discovered Vulcan, although academic incentives being what they are, there were numerous claimed sightings. So Mercury became a "falsifying" observation, but this did not dethrone Newtonian theory, which prospered amidst this and other long known anomalies. Only when Einstein's theory was able to account for this and other anomalies was Newton finally "displaced."

---

[7] Peter Fishburn, *The Foundations of Expected Utility*, Dordrecht, Holland: D. Reidel, 1982. Kenneth MacCrimmon and Stig Larsson, "Utility Theory: Axioms Versus Paradoxes," in Maurice Allais and Ole Hagen, eds., *Expected Utility Hypotheses and the Allais Paradox*, Dordrecht, Holland. D. Reidel, 1979.

But as noted by Imre Lakatos in *The Methodology of Scientific Research Programs* (1978, p. 39) "Einstein's theory is not better than Newton's *because* Newton's theory was 'refuted' but Einstein's was not; there are many known 'anomalies' to Einsteinian theory. Einstein's theory is better than—that is, represents progress compared with—Newton's theory *anno* 1916 (that is, Newton's laws of dynamics, law of gravitation, the known set of initial conditions; 'minus' the list of known anomalies such as Mercury's perihelion) *because* it explained everything that Newton's theory had successfully explained, and it explained also *to some extent* some known anomalies and, in addition, forbade events like transmission of light along straight lines near large masses about which Newton's theory had said nothing but had been permitted by other well-corroborated scientific theories of the day; moreover, *at least some* of the unexpected excess Einsteinian content was in fact corroborated (for instance, by the eclipse experiments)."

It is writ large in the history of science that in scientific communities there *is no falsification in the absence of an empirical content-increasing advance in theory.* There are always, and always will be anomalies, only some of which can be resolved.

Anomalies and their examination do have an extremely important role to play in progressive research programs. But that role is easily obscured by a naive falsificationist rhetorical view to which we all easily fall victim. Our real business is to examine *both* anomaly and nonanomaly in detail—not in just some "stylized fact" form—and take care to reexamine our *ceteris paribus* clauses before claiming to have an anomaly. Then the task is to extend the theoretical framework to (1) account for the anomaly, (2) retain our ability to account for the previous nonanomalies, and (3) permit new testable implications to be derived that were not possible from previous theory. Such constitutes neither refutation nor verification, but progress involving both.

Vernon Smith
Department of Economics
University of Arizona
Tuscon, Arizona

Journal of Economic Perspectives—Volume 2, Number 2—Spring 1988—Pages 185–189

Notes

1989 Nominating Committee of AEA

In accordance with Article IV, Section 2 of the Bylaws of the American Economic Association, President-elect Joseph A. Pechman has appointed a Nominating Committee for 1989 consisting of Alice M. Rivlin, Chair; Vernon J. Dixon, David I Fand, David M. Gordon, Eric A. Hanushek, Irene Lurie, and Janet L. Yellen.

Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

Nominations for AEA Officers: 1989

The Electoral College on March 18, 1988, chose Gerard Debreu as nominee for President-elect of the American Economic Association in the balloting to be held in the autumn of 1988. Other nominees (chosen by the 1988 Nominating Committee) are: Vice-President (two to be elected) Alan S. Blinder, Andrew F. Brimmer, Christopher A. Sims, and Oliver E. Williamson; for members of the Executive Committee (two to be elected) William A. Darity, Jr., Stanley Fischer, John B. Shoven, and Lawrence H. Summers.

In accordance with Article IV, Section 2 of the Bylaws of the American Economic Association, additional candidates may be nominated "by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Assocation for the office of President-elect and not less than 4 percent for each of the other offices. No member of the Association may validly petition for more than one nominee for the Executive Committee, one nominee for Vice-President and one nominee for President-elect." For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

Call for papers. The **Southern Economic Association** will hold its annual meeting at the Hyatt Regency San Antonio, San Antonio, Texas, November 20–22, 1988. The Program Chairman is Dr. V. Kerry Smith, Vanderbilt University. For information on submitting papers, please see the Call for Papers in the October issue of the *Southern Economic Journal*, or write to Dr. Joseph M. Jadlow, College of Business Administration, Oklahoma State University, Stillwater, OK 74078.

Call for papers. The **Association of Managerial Economists** will hold its sixth annual meeting in New York City, December 28–30, 1988. Three sessions of competitively selected papers will be featured in conjunction with the Allied Social Science Association's (ASSA) Annual Meeting. Theoretical, empirical and policy analyses across a broad spectrum of topics will be featured. Both members and nonmembers are invited to submit papers, offer their services as session chairs and discussants, and/or make program suggestions by July 15, 1988, to: Professor Mark Hirschey, AME Program Chair, Jones Graduate School, Rice University, P.O. Box 1892, Houston, Texas 77251-1892. Phone: 713-527-8101, ext. 2549.

Call for papers. The Annual Meeting of the **Association of Environmental and Resource Economists** (AERE) will be held jointly with the AEA in New York, December 28–30, 1988. Those interested in having papers considered for these sessions should send three copies of a one-page abstract to Prof. Charles Howe, Department of Economics 256, University of Colorado, Boulder, CO 80309, no later than June 15, 1988.

Call for papers. ***Studies in Economic Analysis*** is a biannual, refereed, student-edited journal soliciting research articles from both established economists and students. Submission fee is $4.00 for non-subscribers. Submit manuscripts to, or request complete format and style requirements from. The Editors, *Studies in Economic Analysis*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, South Carolina 29208.

Call for papers. The ***Journal of Forensic Economics***. Published twice yearly. Papers dealing with applications of economics in torts and business law. Send inquiries and submissions to: John Ward, Dept. of Economics, University of Missouri-Kansas City, Kansas City, MO 64110-2499. Phone: 816-276-1309.

The **Ferber Award** is given annually to the author of the best article published in the *Journal of Consumer Research*. Articles may be coauthored. For more information, contact Michelle R. Miller, Managing Editor, College of Business Administration, University of Florida, Gainesville, Florida 32611. Phone: 904-392-9837.

Colgate University and Interfaith Hunger Appeal are sponsoring a Faculty Development Workshop on **Issues in International Food Security**, June 13–June 19, 1988. For information contact Ellen Percy Kraly, Project Director, Department of Geography, Colgate University, Hamilton, New York 13346. Phone: 315-824-1000.

The second annual meeting of the **European Society for Population Economics**, ESPE, will be held at the University of Mannheim, Federal Republic of Germany, June 23–25, 1988. Conference registration forms can be obtained from Klaus F. Zimmermann, ESPE-88; University of Mannheim A 5,6; Faculty of Economics and Statistics; D-6800 Mannheim 1; Federal Republic of Germany.

The **International Society for Inventory Research** (ISIR) announces two conferences. The Fourth International Conference on the Foundations and Applications of Utility, Risk and Decision Theory will be held in Budapest, June 6–10, 1988. The Fifth International Symposium on Inventories will be held in Budapest, August 22–26, 1988, in the Hotel Hilton. For information contact Attila Chikán, ISIR, V. Veres Pálné u. 36., Budapest H-1053, Hungary. Phone: 36-1-174-106.

1988–89 Fulbright Grants in Economics. The Council for International Exchange of Scholars (EIES) has announced that a number of 1988-89 Fulbright Grants remain available to U.S. faculty in the field of economics. Scholars in all academic ranks, including emeritus, are eligible to apply and it is expected that applicants will have a Ph.D., college or university teaching experience, and evidence of scholarly productivity. U.S. citizenship is required. In a few countries (of Central and South America and Francophone Africa), knowledge of the host country language is required. Interested scholars are urged to inquire as soon as possible. For information, call or write CIES, Eleven Dupont Circle NW, Suite 300, Washington, D.C. 20036. Phone: 202-939-5401.

The University of Tennessee announces its **International Workshop in Post Keynesian Economics**, "Problems in Income Distribution: Functional, International, Personal," to be held June 26 through July 3, 1988, in the Great Smoky Mountain Resort area of Tennessee. The program is being organized by Professors P. Davidson and J. A. Kregel. For information, contact Louise Davidson, c/o *JPKE*, University of Tennessee, 508 Stokely Management Center, Knoxville, Tennessee 37996-0550.

The **Indian Econometric Society** announces the *Journal of Quantitative Economics* (*JQE*) published biannually by the Delhi School of Economics. The editor is A. L. Nagar. Authors from North and South America who wish to submit their articles to the *JQE* may send them (in triplicate) either to the editorial office of the journal at Delhi School of Economics, University of Delhi, Delhi 110 007, India, or to Damodar Gujarati at the Department of Economics and Finance, Baruch College (CUNY), 17 Lexington Avenue, New York, N.Y. 10010.

The **American Statistical Association** (ASA) and the **Bureau of Labor Statistics** (BLS), under a grant from the National Science Foundation (NSF), are conducting a research fellowship program. Fellows and trainees selected by ASA will work at the BLS for periods of up to one year on specific research activitie relating to BLS programs. For a description of the program and more information, write to Dr. Fred Leone, Executive Director, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402.

The **Department of Economics, University of Illinois at Chicago** has received authorization to grant a Ph.D. in Business Economics. The new degree program is intended to prepare students for research careers in the business community or research/teaching careers in business colleges. Applicants with or without an MBA degree are welcome to apply. The new program complements the department's existing Ph.D. program which emphasizes public policy analysis.

The **American Council of Learned Societies** offers small travel grants to economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship and/or participation. The deadlines for applications to be received are: for meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45th Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Research in the History of Economic Thought and Methodology, published annually by JAI Press, will inaugurate in 1988 a series of archival supplements. The archival volumes will contain hitherto unpublished materials of varying description. Suggestions and contributions may be sent to Warren J. Samuels, Department of Economics, Michigan State University, East Lansing, MI 48824.

The **Economics Department and the School of Natural Resources of the University of Michigan** offer a joint Ph.D. Program in Natural Resource Economics. In addition, both the Economics Department and the School of Natural Resources offer an M.A.; students may pursue a dual degree program combining the two, or a program in one of the Schools. For further information, contact Natural Resources Economics Program, Economics Department, Lorch Hall, The University of Michigan, Ann Arbor, MI 48109, or the Office of Academic Programs, School of Natural Resources, 1024 Dana Building, The University of Michigan, Ann Arbor, MI 48109-1115.

The **National Institute of Mental Health** announces a new research grant program to encourage investigator-initiated research on services to the severe mentally ill including the priority area of financing such services. For information contact: Mr. Paul Widem or Dr. Agnes Rupp, Mental Health Economics Research Program, NIMH, 5600 Fishers Lane, Room 18C-14, Rockville, Maryland 20857. Phone: 301-443-4233.

The *Journal of Futures Markets* offers discount subscriptions to members of the AEA. Members pay $70.00 for 6 issues (regular rate, $142). Outside the U.S. add $37.00 for air service. Members who wish to take advantage of this special rate may send their orders with payment directly to the Periodicals Division, John Wiley & Sons, Inc., P.O. Box 836, Bound Brook, N.J. 08805, and identify themselves as members of the American Economic Association.

Promotions

Robert L. Ayanian: professor of economics, California State University, Fullerton, August 1986.

Colin I. Bradford, Jr.: associate professor in the practice of international management and economics, School of Management, Yale University, July 1, 1987.

Robert W. Campbell: distingushed professor of economics, Indiana University-Bloomington, July 1, 1987.

Robert T. Clair: senior economist, Federal Reserve Bank of Dallas, July 15, 1987.

W. Michael Cox: assistant vice president and senior economist, Federal Reserve Bank of Dallas, July 1, 1987.

Michael D. Duffy: associate professor, Department of Economics, Iowa State University, July 1, 1987.

Jane V. Hall: professor of economics, California State University, Fullerton, August 1987.

Alan J. Harrison: professor of economics, McMaster University, July 1, 1987.

James P. Henderson: professor of economics, Valparaiso University, August, 1987.

Susan A. Hickok: research officer and senior economist, International Research Department, Federal Reserve Bank of N.Y., October 1, 1987.

Charles M. Lucas: senior vice president, International Capital Markets Staff, Federal Reserve Bank of N.Y., October 1, 1987.

Wayne A. Lewchuk: associate professor of economics and labor studies, McMaster University, July 1, 1987.

Robert N. McCauley: research officer and senior economist, International Finance Department, Federal Reserve Bank of N.Y., October 1, 1987.

Christopher McCurdy: assistant vice president, International Capital Markets Staff, Federal Reserve Bank of N.Y., October 1, 1987.

Susan F. Moore: vice president, Statistics Function, Federal Reserve Bank of N.Y., October 1, 1987.

Gerald P. O'Driscoll, Jr.: assistant vice president and senior economist, Federal Reserve Bank of Dallas, July 1, 1987.

Ronnie J. Phillips: associate professor of economics, Colorado State University, July 1, 1987.

Anil K. Puri: professor of economics, California State University, Fullerton, August 1986.

Lawrence J. Radecki: senior research officer, Domestic Research Department, Research Function, Federal Reserve Bank of N.Y., October 1, 1987.

Vincent Reinhart: senior economist, Trade and Commodity Studies Staff, International Research Department, Federal Reserve Bank of N.Y., September 17, 1987.

Charles F. Revier: associate professor of economics, Colorado State University, July 1, 1987.

A. Leslie Robb: professor of economics, McMaster University, July 1, 1987.

Richard C. Schiming: professor of economics, Mankato State University, September, 1987.

John R. Schroeter: associate professor, Economics Department, Iowa State University, August 21, 1987.

David E. Sisk: professor, Department of Economics, San Francisco State University, Fall 1987.

Michael R. Veall: associate professor of economics, McMaster University, July 1, 1987.

Betsy Buttrill White: vice president, Banking Studies and Analysis Function, Federal Reserve Bank of N.Y., October 1, 1987.

David C. Wong: associate professor, California State University, Fullerton, August 1986.

Administrative Appointments

Anthony M. Akel: Dean, Professor, School of Management, The William Paterson College of New Jersey, Summer 1987.

Ana N. Eapen: Chairperson, Department of Economics and Finance, School of Management, The William Paterson College of New Jersey, September 1987.

Warren J. Prunella: Associate Executive Director for Economic Analysis, U.S. Consumer Product Safety Commission, December, 1987.

New Appointments

Linda Bilheimer: senior economist, Mathematica Policy Research, Inc.

Richard Cantor: economist B, Domestic Research Division, Domestic Research Department, Federal Reserve Bank of N.Y., August 3, 1987.

Mark Dynarski: senior economist, Mathematica Policy Research, Inc.

Anne Gordon: economist, Mathematica Policy Research, Inc.

Ohannes Hagopian: assistant professor, Department of Economics and Finance, The William Paterson College of New Jersey, September 1987.

Gikas Hardouvelis: economist B, Domestic Financial Markets Division, Financial Studies Department, Federal Reserve Bank of N.Y., September 2, 1987.

Philip King: assistant professor, Department of Economics, San Francisco State University, Fall 1987.

Donald Mar: assistant professor, Department of Economics, San Francisco State University, Fall 1987.

Dave O'Neill: senior economist, Domestic Research Department, Federal Reserve Bank of N.Y., September 9, 1987.

Lawrence Schwartz, Mathematica Consulting and Research: senior research fellow, Logistics Management Institute, April 1987.

George Sofianos, economist B, Domestic Financial Markets Division, Financial Studies Department, Federal Reserve Bank of N.Y., July 20, 1987.

Sally Stearns: economist, Mathematica Policy Research, Inc.

Vivian Valdmanis: health researcher, Mathematica Policy Research, Inc.

Miscellaneous

C. Phillip Baumel, Department of Economics, Iowa State University: received the Burlington Northern Foundation Faculty Achievement Award in May 1987; 1987 Fellow, American Agricultural Economics Association.

Fritz Efaw: visiting assistant professor, Vanderbilt University, August 1987.

Neil E. Harl, Department of Economics, Iowa State University: was presented the U.S.D.A. Superior Service Award by the Secretary of Agriculture in Washington, D.C., June 1987; received the Henry A. Wallace award at ISU Alumni days in June 1987; received the Discussion Leader of the Year Award for 1986, from the Nebraska Society of Certified Public Accountants.

John Hillard: visiting professor of economics, Vanderbilt University, August 1987.

Wallace E. Huffman, Department of Economics, Iowa State University: received the American Agricultural Economics Association 1987 award for Quality of Research Discovery.

B. F. Kiker: University of South Carolina, Fulbright Scholar, Faculdade de Economia do Porto, Portugal, Spring 1988.

George W. Ladd, Department of Economics, Iowa State University: 1987 Fellow, American Agricultural Economics Association.

Giancarlo Moschini, Department of Economics, Iowa State University: received the American Agricultural Economics Association 1987 Outstanding Ph.D. Dissertation Award.

Dennis Pervis: Economics Department, Iowa State University, Adjunct Associate Professor, Fall 1987.

Dennis R. Starleaf, Department of Economics, Iowa State University: received the Art Johnson Award for excellence in Applied Science.

To Departmental Representatives and Executive Officers

When submitting information for inclusion in "Notes", please observe the following guidelines:

Calls for papers, notice of professional meetings, and other announcements of interest to economists should be submitted in one paragraph that contains all relevant information.

News of individual members should be labeled as to category: (1) deaths; (2) retirements; (3) foreign scholars visiting the US or Canada; (4) promotions; (5) administrative appointments; (6) new appointments; (7) leaves for special appointments; (8) resignations; (9) miscellaneous. Give individual's name, present place of employment, new title, new institution, and date on which change is to occur.

Deadlines for each issue are: Spring, December 15; Summer, March 15; Fall, June 15; Winter, September 15.

Please send all information to the *Journal of Economic Perspectives*, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544. We reserve the right to edit material received.

Journal of Economic Perspectives— Volume 2, Number 3—Summer 1988—Pages 3-23

The Task Force Report: The Reasoning Behind the Recommendations

Bruce Greenwald and Jeremy Stein

The aim of this article is to provide a framework for thinking about the recommendations made by the Presidential Task Force on Market Mechanisms. The Task Force made four broad recommendations:

1) There should be a single regulatory agency (the Federal Reserve was suggested as the leading candidate) to coordinate operating procedures and regulations across the several equity-related markets. The markets for stocks, options and futures are currently subject to different regulatory authorities (e.g., the stock and options markets to the SEC and the futures markets to the CFTC).[1]

2) This agency should supervise the establishment of "circuit-breaker" mechanisms, which would permit an orderly shut-down and subsequent orderly reopening of trading if conditions like those of October 19th and 20th, 1987, were to reoccur.

3) The clearing and settlement systems of the various markets should be unified in order to minimize inter-market credit exposures and the problems that these created on October 19th and 20th.

4) Finally, margin requirements across the different markets should be made consistent.[2]

These recommendations (and the lack of others) are supported by three conclusions that can be drawn from the Task Force's findings.

First, the proper focus of analysis of the events of the October crash should be on "market mechanisms" rather than on fundamental imbalances in the economy as a whole. Analysis should

---

[1] For those readers less familiar with stock market institutions, the Appendix provides a brief introduction to some of the instruments, regulations, and market structures important to understanding the events of mid-October.

[2] The Task Force also recommended that the quality of trade-related data be improved to facilitate future investigations like its own.

■ *Bruce Greenwald is a member of the Technical Staff, Bell Communications Research, Morristown, New Jersey. Jeremy Stein is Assistant Professor in Finance, Harvard Business School, Harvard University, Cambridge, Massachusetts. Both authors were staff economists with the Presidential Task Force on Market Mechanisms.*

focus on the speed and disorder of the crash rather than on reasons for the overall drop in the level of stock prices. None of the Task Force's recommendations, therefore, address widely cited "causes" of the October decline such as the budget and trade deficits.

Second, the instability evident in the events of October 1987 *was not the inexorable limit of a steadily increasing level of day-to-day stock price volatility.* To the extent that recent institutional developments (new investment strategies, derivative instruments, internationalization of markets, increased trading volume, and so on) caused the October crash, they did so by creating the hitherto hidden potential for such a convulsive episode. They did not generate significantly increasing disorder in daily market movements in the period prior to October 1987. Thus, whenever possible, corrective interventions should be designed to address crisis situations with as little effect on day-to-day operations as is feasible. This is the intended goal of the circuit-breakers recommended in the report. The focus on crisis management also underlies the Task Force's preference for a single agency, rather than a committee of regulators, to supervise the markets. In a crisis, a single authority with clearly established responsibility would have an advantage over other coordinating mechanisms. A major benefit to having the Fed perform this role is that it has the resources needed to support crisis intervention as it did on October 20th, when it calmed a growing financial panic by promising to make liquidity available. Finally, the proposals for common clearing and settlement would have little impact on daily market operations but could go a long way towards reducing the likelihood of credit panic during times of extreme price movements.

Third, under the sorts of conditions that prevailed on late Monday and Tuesday, an orderly halt to trading (and subsequent orderly reopening) would have been preferable to what actually took place. Price quotes gave investors virtually no information about either the levels at which their trades would actually be executed or the general state of market demand. Transactions made in these circumstances can hardly be deemed "efficient."[3] Another informational problem concerned rumors about the financial viability of various participants. One important goal of any circuit breaker system should thus be to stimulate information flows during trading halts.

In what follows, we describe how the data collected by the Task Force leads us to the three broad conclusions stated above. Before proceeding, we should emphasize the consciously restricted scope of this article. Our descriptive coverage of the actual crash is limited to the aspects that most clearly distinguish it as a unique crisis event: the breakdown of price setting mechanisms on late Monday and early Tuesday, and the credit related problems and ensuing near-panic that arose on Tuesday. These aspects are crucial to an understanding of the Task Force's recommendations. They do not, however, answer the question of what external pressures led to the crisis: Who was doing all the selling? Who was failing to buy? We would also like to stress that the discussion which follows does not represent any kind of official Task Force position.

---

[3] Economists accustomed to free trade arguments might have difficulty believing that closing markets can ever be a good thing. After all, trading is always voluntary; so how can removing this freedom make anyone better off? We address this question later in the article.

Rather, it embodies our view of how some major aspects of the Task Force's findings should be interpreted, which is a very different matter indeed.

Market Mechanisms versus Fundamental Economic Factors

A necessary first step in separating these two categories of possible causes of the October market decline is to define the basic distinction involved. Fundamental economic factors are those forces which affect the future returns (both real and nominal) paid on the securities which are traded in financial markets. These include future interest rates, exchange rates, price levels, profits and taxes. By extension, they include factors like the budget and trade deficits and increases in corporate and private debt which are likely to affect future returns. In theory, these are the forces which should govern the level of stock prices. Changes in stock prices should, therefore, be attributable to new information about these factors. In practice, such new information appears historically to have arrived in relatively limited amounts over long periods of time. Thus, fundamental economic factors could be responsible for a substantial decline in the level of stock prices but could not, by themselves, explain why the decline should be as sharp and disorderly as that of October 1987.

What we refer to here as market mechanisms are factors associated with the interaction of investors in the various equity-related marketplaces. Among these market mechanisms are portfolio insurance and other trading strategies, market-making systems, and linkages between the various stock exchanges and the futures and options markets. They offer more promise as explanations for the unprecedented suddenness of the October market move and the consequent dislocation of financial markets.[4] These market mechanisms were, for several reasons, the proper focus of the Task Force's investigation.

The first and most important reason for not identifying and evaluating fundamental causes of the October events is that the record on the long-run magnitude of the current stock market decline is far from complete. The Dow's three month fall of 30.5 percent from the market peak of 2722 on August 25 to the level of 1892 on November 19 is smaller than many postwar declines and is dwarfed by the decline of 89 percent from the 1929 peak to the 1932 low which coincided with the start of the Great Depression. If the market stabilizes at its current level, then the long-run magnitude of the recent crash will, from the perspective of history, have scarcely justified special attention. If, on the other hand, the market continues along a path similar to that experienced in 1930–32, it will, regardless of the sharp nature of the October drop, justifiably be the subject of intense study. Unfortunately, we do not yet know which of these possibilities will occur.[5]

[4] A third category of factors relate to investor psychology. However, for the purposes of analysis these factors can be divided into changes in investors' outlooks which change relatively slowly over time (although the magnitudes of such changes may be substantial) and changes in the attitudes of active traders which occur rapidly in response to changing market conditions. We will treat the former as fundamental economic factors and the latter as market mechanisms.

[5] Thus far, however, we are behind the pace set by the 1929–1932 decline. The market fell 34.8 percent from its peak in September 1929 to its year-end close, and another 33.8 percent in 1930, so that it stood at only 43.2 percent of the September 1929 peak by December 1930.

Table 1
Substantial Market Declines and Subsequent Economic Impacts

Dates	*Percent change in DJIA*	*Initial DJIA*	*Percent GNP change (following 12 months)*[a]	*Recession*
a. United States (Post-War):				
5/29/46 to 10/9/46......	(23.2)	212.5	6.4[b]	No
12/13/61 to 6/26/62	(27.1)	734.9	4.9	No
2/9/66 to 10/7/66.......	(25.2)	995.1	2.4	No
12/3/68 to 5/26/70......	(35.9)	985.2	(1.3)[c]	Yes
8/22/73 to 11/5/84......	(44.4)	1,051.7	(3.6)	Yes
9/21/76 to 2/28/78......	(26.9)	1,014.8	3.5	No
4/27/81 to 8/12/82......	(24.1)	1,024.0	1.2	Yes
8/25/87 to 11/19/87	(30.5)	2,722.4	N/A	N/A

Dates	*Percent change in FT30*[d]	*Initial FT30*[d]	*Percent GNP change (following 12 months)*[a]	*Recession*
b. United Kingdom:				
6/30/55 to 2/28/58......	(29.8)	39.97	2.4[b]	No
4/30/61 to 7/31/62......	(21.2)	64.05	1.1[b]	No
1/31/69 to 6/30/70......	(31.5)	107.62	2.7	No
12/31/72 to 12/31/74...	(69.5)	124.83	0.2	Yes

Dates	*Percent change in TSEI*[e]	*Initial TSEI*[e]	*Percent GNP change (following 12 months)*[a]	*Recession*
c. Japan:				
9/30/49 to 6/30/50......	(46.0)	8.46	58.6[b]	No
2/28/53 to 4/30/53......	(24.4)	21.97	21.1[b]	No
7/31/61 to 10/31/61	(37.7)	68.40	5.8[b]	Slowdown
1/31/73 to 10/31/74	(36.6)	145.87	(2.7)	Yes

[a] From midpoint of decline.
[b] Industrial production (real GNP figures not available quarterly).
[c] Changes for one quarter backshift in window.
[d] Financial Times 30 stock index.
[e] Tokyo Stock Exchange Index.
Sources: Salomon Brothers Research: "OECD Economic Statistics; International Financial Statistics; U.S. Economic Report of the President."

Second, large fluctuations in stock prices with no clear fundamental explanation (either prospectively or retrospectively) have historically occurred with some regularity both in the United States and abroad. Table 1 documents the fact that substantial stock market declines are often not followed by noteworthy downturns in the economy. Since World War II significant falls in the United States stock market have been followed by recessions slightly less than one-half of the time. In the United Kingdom and Japan a similar ratio prevails. In the most striking instance, a 70 percent decline in the British stock market between December 1972 and December 1974 was followed

Table 2
Stock Market Performance in October 1987
Versus Underlying Economic Conditions–International Comparisons
[in percent]

Country	*October price decline*	*P / E October 1987*	*Long term Govern-ment rate*	*Rate of inflation*	*Unemploy-ment rate*	*Growth rate*[a]	*Trade deficit*[b]	*Govern-ment deficit*[b]
United States......	21.5	18.9	9.42	4.1	6.9	5.2	(3.3)	(5.3)
Australia	44.7	19.2	13.25	8.1	N/A	N/A	(1.1)	(1.0)
Canada	22.2	26.1	10.44	4.2	9.4	4.4	1.4	(4.2)
United Kingdom..	21.7	16.0	9.92	4.0	11.6	3.8	0.1	(2.2)
France	18.6	14.9	9.85	3.3	10.7	1.1	1.0	(2.6)
Germany	17.7	15.4	6.20	0.6	7.5	2.2	5.8	(1.5)
Italy	12.3	17.0	10.58	4.3	6.1	3.3	0.5	(12.2)
Japan	7.5	61.7	4.44	0.2	2.9	5.0	4.4	(4.9)[c]

[a] Industrial production change (October 1986 to October 1987).
[b] Percent of GNP.
[c] Calculated from the increase in net government debt outstanding (typically understates deficit which is not reported to the International Monetary Fund).
Sources: Morgan Stanley/Guardian International Statistics; "International Financial Statistics," U.S. Economic Report of the President.

by only a mild recession in 1974–75. Over the longer term, there have been large secular changes in both the dividend yields and price earnings ratios of common stocks that have been largely unrelated to changes in either interest rates or future earnings.[6]

Third, the difficulties associated with identifying fundamental causes are underscored by the international nature of the October decline in the market. Economies as diverse as those of the United States, the United Kingdom, Germany, France, Italy and Australia all experienced stock market declines of comparable magnitudes, as shown in Table 2. At the same time, indicators of potential and current economic problems differ widely among these countries. For example, the U.S. federal government deficit is frequently cited as the root cause of the October stock market decline. For 1986 this was equal to 5.3 percent of GNP and during October 1987 the United States stock market declined by 22 percent. The comparable Australian government deficit was only 1.0 percent of GNP, but Australian stock prices declined by 45 percent in October 1987. The Italian deficit was more than 12 percent of GNP and yet Italian stock prices declined by only 12 percent.

Fourth, even if it were known with certainty that the October market decline was driven by fundamental factors, it is unrealistic to expect the Task Force to make reasonable policy recommendations in these areas within its two month reporting period. For example, despite extended study of the effect of government budget deficits, there is not yet agreement on how they should be measured, or on the channels through which their effects are transmitted. Correspondingly, in long-standing discussions of the impact of "liquidity" on financial markets, there is equally little agreement on how liquidity should be quantified or exactly how it influences stock prices. It is difficult, therefore, to see how the Task Force could expect to sensibly apportion responsibility for the October events to causes like the budget deficit and "liquidity." It is even more difficult to see how reliable policy prescriptions could be provided based on how these factors operate.

---

[6] See Campbell and Shiller (1987) as one example from a large and growing literature.

Finally, the Task Force on Market Mechanisms, as both its name and its limited reporting time suggest, was created in response to the extraordinary events that occurred on October 19 and the following days. What distinguished these events was the unprecedented rapidity with which prices fell, the huge volume of trading and the consequent dislocation in financial markets. Thus, whatever the causes of the original downward pressure on the market, the clearly implied mandate of the Task Force was to focus on those factors which transformed this downward pressure into the alarming events of these critical days and to recommend measures to ensure, as far as possible, that future market fluctuations do not take on the extreme and potentially destructive character witnessed in October 1987.

Any contribution of fundamental causes to the recent market decline should not, of course, be ignored. To the extent that existing imbalances in the budget, foreign transactions, savings, corporate asset positions and other fundamental factors are perceived to be problems they merit study. A heightened focus of attention on these subjects may represent one of the few desirable side benefits of the October market decline. Appointment of a group with more time and greater resources to investigate these questions thoroughly before action is taken might be appropriate. This Task Force was not, however, equipped to deal with such issues in a useful way.

Historical Evidence on Stock Market Volatility: The Case Against Structural Reforms

The past several years have seen a variety of changes in equity market institutions: pension funds and mutual funds have grown in importance; trading volume has increased dramatically; index future and options contracts have emerged as widely used tools for speculation and hedging; and foreign markets have risen in relative stature.[7]

Some observers have suggested that one or more of these trends has led to steadily increasing stock market volatility. In this view, the events of October 19th and the surrounding days are simply the most visible symptoms of an inexorable movement towards greater amplitude in price changes.

If this position is accepted, it implies that "structural reforms" might be warranted as a method for preventing another crash. By "structural reforms" we mean

[7] Some examples of the above trends: In the second quarter of 1987, pension and mutual funds together controlled 26.2 percent of U.S. equity, up from 20.2 percent in 1981. The average stock on the NYSE was traded .64 times in 1986, triple the .21 turnover of 1977. And in 1986, approximately 20 million S&P 500 futures contracts (representing a dollar value of around $2.5 trillion, roughly the same as that traded on all U.S. equity markets) were traded. These contracts were nonexistent six years earlier.

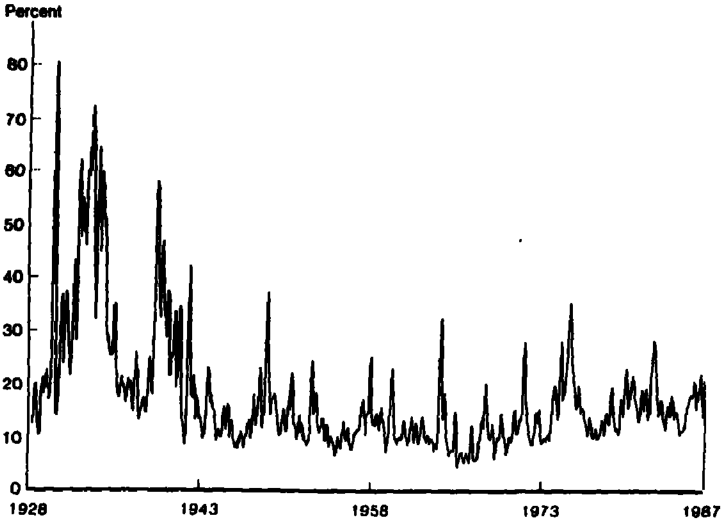


Fig. 1. 60 day historical volatility S&P 500 Index

policies that alter the behavior of the markets on a day-in, day-out basis, presumably reversing some of the above-noted trends. A couple of examples that spring to mind are the abolition of index futures and options or a trading tax designed to lower stock turnover.[8]

We find little evidence, however, to support the notion that institutional changes have led to significant increases in daily volatility. We are left to conclude that the October crash represented a fundamentally unique event, not easily extrapolated from recent trends in market participation and volatility. This casts doubt on the need for continual interference in the market as a requisite for disaster prevention.

For the most part, we focus on U.S. stock price data. Figures 1 and 2 are based on daily prices from the Standard and Poor's 500 index from 1928 through September 1987. We also touch briefly on international evidence, in order to see how trends in volatility in Japan and Germany compare to those in the United States.

Figure 1 displays one common measure of volatility: the annualized standard deviation of daily percentage returns, calculated using the preceding 60 trading days.[9]

---

[8] Although we will argue below that such measures are not the most effective way to prevent a recurrence of a market crisis, this should not be taken to mean that there are no grounds for considering them. For example, Summers (1987) puts forward other arguments for a trading tax. One "structural reform" that we do regard as indefensible on any grounds is the abolition of index arbitrage *without* any curbing of futures trading. If futures markets are to exist, one has to accept as a fact of life that price pressures originating there will be transmitted to the underlying cash markets. Index arbitrage is simply an efficient mechanism for performing this transmission, and abolishing it will not allow us to live in a fantasy world where massive futures activity can take place without affecting stock prices.

[9] Volatility at a given point in time is calculated as follows: Take the preceding 60 trading days worth of daily data. Calculate the standard deviation of percentage price changes, using those 60 observations. Finally, annualize by multiplying by the square root of 250 (since there are about 250 trading days in a year).

500 futures market was open throughout October 19 and October 20, except between 12:15 p.m. and 1:05 p.m. on October 20. Only the options markets were closed for substantial fractions of October 19 and 20.[10]

However, availability and trading volumes are not the sole measures of market performance. Effective market mechanisms manage trades at prices that are well-defined for market participants and that efficiently convey information about current market conditions. In neither of these respects did equity markets perform adequately on October 19 and 20.

Broadly speaking, stock prices (as measured by the Dow Jones index) fell from a level of 2500 at the close on October 13 to a level of between 1900 and 2000 once markets stabilized after October 23. From October 13 to October 16, approximately half of this decline took place in a reasonably orderly way. Intra-day upward movements in market indices against the dominant downward trend never exceeded 2 percent and maximum spreads between futures and cash index prices exceeded 2 percent only slightly and briefly (see Figure 3 for a plot of prices October 14-20).

At the opening of trading on October 19, this performance began to deteriorate. Delayed openings in the stocks of many large companies created uncertainties about the true level of the cash indices. An apparently substantial discount between the futures price and cash index levels (computed using October 16 closing prices for unopened stocks) led to significant cash sales and futures purchases by index arbitrageurs which accounted for roughly 15 percent of NYSE volume between 9:30 a.m. and 10:30 a.m. In retrospect the discount was illusory.[11] At the same time, NYSE specialists appear to have substantially misestimated appropriate opening prices for several individual stocks.[12] For example, two large capitalization NYSE-listed stocks which did not begin trading until about 10:30 a.m. on October 19, opened down 17 percent and 19 percent. In the next hour, during which the Dow fell by 1.4 percent, these two stocks rose by 13 percent and 16 percent respectively, recovering roughly 80 percent of their opening losses. Individuals who sold at the open, thus suffered more than 10 percent losses compared to those whose sell orders were executed just one hour

---

[10] Trading delays on the NYSE arise when the specialist faces an imbalance of buy and sell orders that he feels is beyond his obligation to resolve. In such cases, he can request permission from a floor official to halt trading, during which time he displays price indications in an attempt to reduce the imbalance. The S&P futures trading halt was ordered by the CME. In the options market, each option series is opened separately in a "rotation," which ordinarily takes about 20 minutes. After this rotation, all the options trade freely. On October 19th and 20th, however, these rotations took several hours, resulting in a severe curtailment of free trading.

[11] After correcting for the fact that October 16 closes overstated the prices of unopened stocks and that unopened stocks suffered greater opening declines than those which began trading shortly after 9:30 a.m., the cash levels of the major stock indices closely tracked their futures prices throughout the morning of October 19. This is confirmed by arbitrageurs who report earning no profits from their morning trades.

[12] A specialist faced at the opening on October 19 with a large excess of "at market" sell orders over buy orders had great discretion in determining an opening price at which he was willing to cover the resulting imbalance from his own capital. Consider, for example, the specialist (with only a smattering of limit buy orders) who faced 400,000 shares in at market sell orders and only 350,000 shares in buy orders. He would purchase the balance of 50,000 shares of his own account, but might do so at a price anywhere from 5 percent to 15 percent down from the previous closing price.

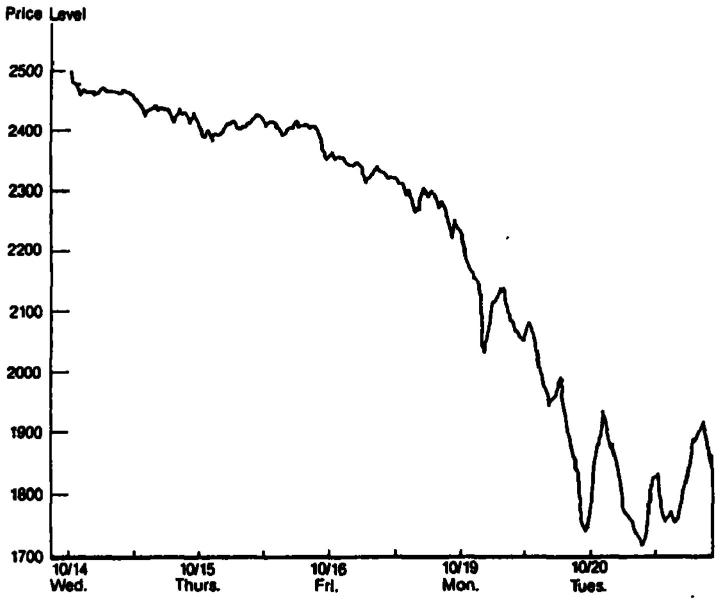


Fig. 3. Dow Jones Industrial one minute chart. October 14, 1987 to October 20, 1987

later. Nevertheless, on the morning of October 19 these cases were the exception rather than the rule (see Table 4) and by early afternoon the market appeared to have stabilized at a Dow level in the 1900–2000 range.[13]

However, from the mid-afternoon on October 19 until the late afternoon on October 20, equity markets became separated from any stable economic reality. In the cash markets, this did not occur until about 3:00 p.m. on October 19. In the final hour of trading on October 19, the Dow, which had reached a level of 1950, fell by 220 points or 11.2 percent. On October 20, the Dow opened up 12.1 percent from the October 19 close, eliminating the loss which had occurred in the final hour of trading on October 19.[14] From these opening levels stock prices fell immediately and rapidly. By noon on October 20, the Dow had fallen 11.4 percent, almost exactly to the level of the close on October 19. Finally, when the Dow subsequently stabilized, it recovered most, if not all, of this loss.[15] Figure 3 illustrates these whipsaw movements.

[13] Although we are describing the events of October 19th and 20th in terms of the popular Dow Jones index, it should be noted that the broader S&P 500 index tells essentially the same story.

[14] The openings of the Dow stocks were fairly evenly spread out between 9:30 a.m. (the official start of trading) and 11:30 a.m.

[15] To put these "whipsaws" in perspective, note that each leg is roughly comparable in magnitude to the previous record one day fall of 12.8 percent on October 28, 1929. Even more than the fact that stocks lost 22.6 percent of their value overnight, these whipsaws seem impossible to rationalize as driven by news about "fundamentals." In this sense, they would appear to go a long way towards discrediting more extreme versions of the "efficient markets" hypothesis.

Table 4
A Sample of NYSE Price Changes October 19 and 20
(Percentage price change)

Stock no.	*Close October 16 to open October 19*	*Open October 19, to 11:30 am*	*3 pm to 4 pm, October 19*	*Close October 19 to open October 20*	*Open October 20 to 11:30 am*
1	(7.63)	1.30	(15.69)	6.67	(11.98)
2	(4.04)	1.05	(4.87)	0.61	(7.23)
3	(10.49)	8.63	(13.57)	31.15	(15.00)
4	(7.07)	(0.29)	(8.77)	12.68	(16.09)
5	(9.09)	4.17	(8.70)	(6.83)	N/A
6	(16.71)	12.50	(20.81)	(17.30)	N/A
7	(8.05)	0.63	N/A	19.40	(6.56)
8	(19.04)	16.07	(6.49)	11.05	(9.41)
9	(3.86)	3.24	(14.23)	8.33	(16.34)
10	(8.15)	4.13	(11.64)	16.22	N/A
11	(10.38)	4.27	(9.06)	10.70	(23.67)
12	(5.06)	0.67	(6.69)	15.50	(13.42)
13	(10.99)	7.10	(6.79)	1.88	N/A
14	(9.54)	1.17	(11.86)	19.23	(27.42)
15	(12.30)	9.17	(4.19)	7.80	(18.42)
16	(4.39)	1.83	(14.00)	13.18	(7.19)
17	(3.53)	(0.46)	16.72	27.09	(15.38)
18	(10.36)	4.08	(8.15)	22.58	N/A
19	(5.81)	0.44	(7.25)	20.64	(8.77)
20	(5.15)	0.00	(19.81)	24.42	(13.55)
21	(9.27)	(3.52)	(13.18)	19.44	(1.52)
22	(8.93)	2.94	(13.45)	0.00	(7.42)
23	(4.55)	(4.76)	(20.75)	15.33	(14.29)
24	(10.92)	(0.47)	(12.04)	8.24	(14.67)
25	(6.98)	(1.25)	(4.64)	(6.99)	(11.30)
26	(4.08)	(8.50)	(10.00)	(7.79)	2.89
27	(7.16)	6.79	(10.63)	22.32	(16.85)
28	(6.76)	(3.94)	(11.11)	2.56	(7.86)
29	(12.33)	10.16	(3.88)	1.54	(14.39)
30	(4.13)	(1.41)	(4.26)	(10.00)	N/A
31	(4.37)	(3.76)	(8.50)	(1.83)	(11.80)
32	(15.70)	9.80	(20.04)	16.77	N/A
33	(2.96)	(1.51)	(5.23)	(1.38)	(1.75)

"N/A" means the stock was not open at the relevant time.

Gyrations in the futures market during this time were similar to fluctuations in the cash markets, but more extreme. In the final two hours of trading on October 19, the S&P 500 futures contract fell 17 percent. Within minutes of the opening on October 20, the contract had recovered this loss, rising 17 percent. During the next two hours the S&P 500 futures contract fell by 25 percent until trading was halted at 12:15 p.m. (in the same period, the S&P 500 index itself declined by only 12 percent). Yet, as in the case of the cash market, this entire loss was recovered as the futures market stabilized on subsequent trading days.

The price behavior of individual stocks after 2:00 p.m. on October 19 mirrored the disorder in the aggregate indices. Declines of 20 percent or more in the final hour of trading on October 19 followed by opening price increases on October 20 that more than made up these losses were commonplace, as shown in Table 4. These opening increases on October 20 then tended to evaporate rapidly in the immediately following hours of trading, although they were subsequently made up again as stock prices stabilized later in the week. In the over-the-counter market, this disorder was reflected in differences of 20 percent or more between the prices at which essentially contemporaneous trades were executed.

Trading under these conditions has three unfortunate economic consequences. First, investors cannot know with any precision at what prices their orders will be executed.[16] Investors who bought at the open on October 20 often paid as much as 20 percent more than either investors whose trades were executed an hour later or than the October 19 closing prices on which many purchase decisions were based. In OTC stocks, comparable price uncertainties existed throughout October 19 and 20.

Second, the information conveyed by prices may be confusing or actively misleading. Many NYSE specialists appear to have been overly optimistic in setting opening prices on October 20. The rapid run-up in futures prices may have been based on these opening prices and the resulting inferences concerning equivalent or greater demands for as yet unopened stocks. However, once the opening gains began to reverse themselves rapidly after the openings, investors outside the NYSE must have been highly uncertain of the true state of market demand. Whether or not the optimistic October 20 opening prices were due to conscious manipulation by NYSE specialists, confidence in the reliability of price-based inferences about the state of market demand for NYSE stocks must have evaporated with the opening gains. Under these conditions, it is not surprising that panic began to drive prices below what many investors would otherwise consider "reasonable" long run levels.[17]

Third, since prices form the basis for investor and lender calculations of net worth and solvency positions, the absence of reliable price data worsens uncertainties in these areas and can lead to unnecessarily large reductions in the credit available to market participants. Without such credit, market-making activities must be curtailed, which then further intensifies price volatility.

On October 19 and 20, these potential solvency and credit concerns were exacerbated in two important ways by the existence of several distinct clearing and settlement systems for futures, options and stocks.

---

[16] This is literally true for "at market" orders. Of course, by submitting a "limit" order (i.e. an order to buy only at prices below a certain limit) an investor can protect himself from buying at an unexpectedly high price, or selling at an unexpectedly low one. However, limit orders do not represent an especially attractive alternative under the conditions of October 19th and 20th. An investor's threshold price should depend on his most current information, which includes the current market price. Under very volatile conditions, this can mean resubmitting limit orders on an almost continuous basis, which would have been extremely difficult to accomplish.

[17] The impact of extremely rapid changes in futures and cash prices on investors without access to fully current information would have been similar. These investors would be unable to infer reliably either the existing terms of trade or the state of demand from price quotes available to them.

First, the existence of separate clearinghouses for stocks, futures and options means that uncertainties about the status of trades (for example, whether the other side of a trade would be solvent) applied to the net positions of trading firms in each clearinghouse separately. Thus, if a particular firm had hedged its futures trades (or options positions) with cash stock positions, the margin of safety implicit in these positions would not be apparent to the individual clearinghouses and their members. The futures clearinghouse would see only the futures side of the overall transaction and would have to infer overall exposure from that position alone. Under these circumstances, uncertainties about firm solvencies and the ultimate integrity of the settlements process created by rapidly fluctuating prices would be larger than they would be if a common clearing system took into account only *net* trading positions.

Second, during October 19 and 20, each of the individual clearinghouses placed its own separate and substantial intra-day margin demands on member firms. Margin calls to member firms on October 19 from the CME (futures) and OCC (options) clearinghouses, which had largely overlapping memberships, amounted to $3 billion. Margins calls on the morning of October 20 were a further $2 billion. Moreover, in the pervasive disorder of the crash, payments from the clearinghouses to members were often significantly delayed.[18] Clearinghouse member firms (large investment banks and brokerages) were forced to borrow substantial short-term funds to cover these demands at a time of grave uncertainty over their creditworthiness. Without the timely intervention of the Federal Reserve System the resulting intensification of credit problems may have had very serious consequences, including perhaps the failure of one or more major investment banks. Since the positions involved were largely offsetting, a common clearing system would have greatly reduced the demand for margin funds and the consequent strain on the credit system.

Conclusions and Extensions

The Crisis Nature of the Crash

The discussion has repeatedly stressed a simple theme: the stock market crash of October 1987 was a fundamentally unique event, wholly different in character from normal market behavior. This point was first made in context of historical volatility data. Prior to the crash, there was no significant upwards trend in volatility from which one could reasonably have extrapolated the events of October 19th and 20th.

The same point reemerged when we looked at market performance during the crash, particularly on late Monday and on Tuesday. The usual informational function of prices was destroyed; trades consummated only minutes apart were executed at wildly different prices, so that an investor submitting a market order had virtually no idea where it would be completed. The magnitude and uncertainty of price movements was also more than the existing clearing mechanisms were prepared to handle: the potentially huge and unmeasurable credit expenses that developed almost instantaneously led to dangerous rumors and a near financial panic. Although total disaster was averted in this regard, solvency concerns certainly did impair the market's performance. In just one example, many traders reported that they avoided the futures market altogether, for fear that the clearinghouses would fail.

---

[18] For example, $1.5 billion in payments to two clearinghouse members on October 20, which were due before 10:00 a.m., did not arrive until 3:00 p.m.

Crises such as the one in October remind us that we do not live in the economist's idealized world of "frictionless markets." The stock market's infrastructure—including computerized trading systems, and market making and clearing mechanisms—are only equipped to handle a limited rate of volume flow and price change. These "institutional constraints" are usually fairly invisible; but when the pressure on them is great enough, a variety of breakdowns can occur.

In a world without institutional constraints, a one-day 500 point stock market fall would look just like a scaled-up version of a one-day 50 point drop, or like 10 days of 50 point declines strung together. In such a world, the laissez-faire logic of "letting prices equilibrate as rapidly as possible without interference" would be compelling. However, since this is clearly not the world we live in, it is difficult to use laissez faire logic to argue against well-planned and well-timed interventions during moments of market crisis. The case for pre-planned circuit breakers is further strengthened if it is noted that institutional constraints act as ad hoc circuit breakers anyway when the pressure is great enough. The October crash saw long backups of order flows and trading halts in numerous stocks, options and futures; and there were some less than responsive specialists and market makers throughout the system.

The Rationale for "Informative" Trading Halts

Up to this point we have been deliberately vague about what "circuit breakers" really are because the language of the Task Force Report was also vague in this regard. Now, however, we will attempt to push our interpretation of the October crash a little further and see where this leads in terms of specific recommendations. We will also try to address some of the objections that are likely to be raised by such recommendations.

Circuit breakers have been discussed in terms of both price limits and trading halts. Our analysis has suggested that one of the strongest arguments for a trading halt was the breakdown of normal information transmission that occurred on late Monday, October 19, and on Tuesday, October 20. The large price movements of the previous Wednesday, Thursday and Friday were not in themselves a terrible problem, as information still flowed smoothly and participants were generally aware of their trading opportunities.

We see nothing inherently wrong with large price movements, as long as they are "fair"—that is, as long as there are not tremendous asymmetries of information between, for example, specialists and their customers. Thus, when information transmission breaks down, the primary function of a circuit breaker should be to reinform participants. This could be accomplished by an orderly trading halt followed by an "open order" period of fixed duration. During the "open order" period, specialists' books of both limit and market orders would be open for general inspection.[19] Orders entered during this time could be changed or withdrawn, since no trades would be executed. At the end of the "open order" period an initial price would be set and trades would be executed, with orders entered during the period receiving priority in execution over subsequent orders (to provide investors with some incentives to enter their orders). Once this has been done, the markets can be reopened, and prices can be allowed to adjust freely.

---

[19] A somewhat stronger recommendation, but a potentially quite sensible one, would be to have specialists' books opened regularly (say, every day at the end of trading) regardless of whether market conditions were chaotic or not.

This informational view of circuit breakers respects many of the criticisms that have been raised regarding market intervention. We do not pretend that closing trading will prevent a needed price adjustment from happening; the only aim is to have this price adjustment carried out under as fair circumstances as possible. Also, we recognize that trading halts may prevent some investors from making the sales necessary to meet margin calls, thereby exacerbating credit related problems. However, it should be pointed out that many of the credit related problems seen in the October crash were not due to actual payments delays, but rather to unfounded rumors about financial solvency. If trading halts serve to improve the information of market participants, they may be helpful in quashing such rumors and the accompanying difficulties.

If trading halts are *not* carefully designed to disseminate information and to produce, within a reasonable time, a new price at which trading can resume, they may wind up being worse than the disease they seek to cure. Thus, the Hong Kong market shutdown was, in the absence of a reopening mechanism, both a disaster and not an appropriate measure of the efficacy of a properly designed circuit breaker system. And given the lack of advance planning, the choice facing U.S. policy makers who considered trading halts on October 19th and 20th was not an enviable one.

Many economists will have a natural antipathy towards trading halts for two reasons. First, there is the free trade argument: aren't people better off having an opportunity to transact, even under chaotic circumstances? They do so freely, and can always not trade if they don't like the risks inherent in doing so. Second, won't more information be revealed through continued trading than during a shutdown?

In our view, these objections overlook a very basic and important externality effect. When people are attempting to draw inferences from market prices, those prices are in a sense a public good—the more informative they are, the better. An individual trader does not take into account the fact that his (rational) actions may contribute to the overall disorder in such a way as to make prices less informative to others. When these externalities are significant in magnitude (as it is likely they were at times during October 19th and 20th), intervention that seeks to disseminate information and establish prices based on broader consensus can make everybody better off.[20]

[20] A formal model along these lines is developed in Stein (1987). There it is shown how the rational, profit-maximizing actions of one group of traders can lower the informativeness of prices to other traders and hence reduce social welfare. (See Greenwald and Stiglitz (1986) for a more general treatment of informational externalities.) Empirically, the work of French and Roll (1986) is also suggestive. They find that the variance of prices is greater over several day periods when the market is open every day, as compared to otherwise equivalent periods when the market is closed on some days. In other words, even under "normal conditions," trading appears to generate additional volatility beyond that which can be attributed simply to the arrival of news.

A corollary to this externality effect is that free trading may not always be the best way to encourage traders to bring their demands to market. If prices become very uninformative, the attendant risks may discourage many people from attempting to trade. The purpose of an open order period would be to counter this problem. By allowing traders to look at the demands and supplies of others before committing to a decision, an open order period could reduce the risks arising from poor information and hence provide incentives for "value buyers" to make their liquidity available to the market.

More generally, episodes like the one on October 19th and 20th can have adverse externality effects on a much broader group of people than simply those wishing to trade during that period. Equity markets are a critical source of capital for firms, and convulsive incidents such as that of the past October may, by increasing investors' perceptions of risk, have a long-lasting impact on the cost of equity capital. Furthermore, stock holdings represent a significant fraction of household wealth, and stock prices are a widely followed indicator of economic health. The uncertainties about household wealth and the future direction of the economy that are created by violent stock market swings could conceivably lead to reductions or deferrals of consumption and investment spending and hence to a decline in overall economic activity. Thus, if trading halts can temper the nature of such swings, the benefits could potentially be widespread.

A Final Suggestion: Contingent Capital for Specialists

Although the Task Force report did not make any specific recommendations concerning the capitalization of specialists, it did point out the subject as worthy of further consideration.

Many observers have suggested that the capital base of the specialists should be augmented so that they are better able to stabilize markets in the face of the type of selling pressure seen in October. Our view of the crash as a rare event rather than as a part of an evolving trend hints at another approach. Maintaining a significantly larger amount of capital on hand on a daily basis is an expensive proposition; moreover, it would be unnecessary 99 percent of the time. What might be more effective is some sort of contingency scheme wherein the exchange as a whole would (from specialists' contributions) buy insurance from a consortium of insurance companies. In return for these premiums, the insurance consortium would agree that, in the event of an overall market move in excess of some fixed limit (perhaps 5 percent) on volume exceeding another limit (perhaps more than 135 percent above "normal"), it would buy some fraction (perhaps 80 percent) of the closing inventories of the specialists at cost (appropriately defined). This kind of term insurance would obviate the need for idle capital, minimize the incentive problems that arise when specialists are themselves required to absorb fully any losses associated with resisting selling pressure like that which materialized on October 19 and 20, and add only about 1 to 2 percent to the existing equity holdings of the insurance companies.

It should, however, be clear that this scheme, like the specific circuit breaker mechanism described above, is one with many ramifications that require careful and detailed evaluation. Since no such evaluation was possible in the time available to the

Task Force, the actual recommendations of the Task Force quite properly focus on general areas of concern rather than specific recommendations.

Appendix
Introduction to Stock Market Institutions

This appendix is designed to serve as a brief introductory guide for readers less familiar with the instruments, marketplaces and trading strategies that are important for understanding the events of mid-October.

Stocks, Futures Contracts and Options Contracts

Shares of stock are claims of ownership on corporations. The price of a stock in an effectively operating stock market will depend largely on the current performance and future earnings prospects of a corporation. Futures contracts and options contracts are not corporate ownership claims. They are "derivative" instruments whose value depends primarily on the underlying price of the stock or portfolio of stocks from which they are derived. A futures contract, for example, is an agreement to buy or sell a specified amount of some item at a specified price at a specified time in the future; the value of the contract will vary according to how the present price compares to the price specified in the contract. The most heavily traded equity-related futures and options contracts are based upon certain standardized portfolios of stock such as the Standard and Poor's 500 Stock Index ("S&P 500"), the Standard and Poor's 100 Stock Index ("S&P 100") and the Major Market Index of 20 stocks ("MMI").

Exchanges and Market Making

Stocks are traded on the New York Stock Exchange (NYSE) and American Stock Exchange (Amex), as well as on several other exchanges throughout the country. Other stocks are traded in the over-the-counter (OTC) market, a dealer market connected by computers and telephones.

The S&P 500 futures contract is traded on the Chicago Mercantile Exchange (CME), and the MMI futures contract is traded on the Chicago Board of Trade (CBOT). The preponderance of the daily volume of index futures trading takes place on the CME. Although the value of open interest in the futures contracts is only a small fraction of th value of NYSE stocks, the value of the stocks represented by the volume of futures contracts traded on the CME daily is typically about twice the value of stocks traded on the NYSE daily.

Options contracts on the S&P 100 are traded on the Chicago Board Options Exchange (CBOE). The Amex trades an option on the MMI. Options whose value is related to individual stocks are also traded on various exchanges.

A "specialist" system is used by the various stock exchanges for exchange-listed stocks. Under the specialist system, a single dealer is given the right to make the market in a specific stock or option on the exchange. In return, the specialist assumes the responsibility to make an "orderly" market by buying and selling from inventory. In the competitive market maker system, competing dealers set the price of an options or futures contract in an auction process. A competitive market maker system is used by the CBOE for options, and by the CME and the CBOT for futures. The OTC also uses a competing dealer system to make markets. A hybrid system employing both specialists and competing market makers is used for options sponsored by the stock exchanges.

Regulation

The stock, futures and options exchanges organize, manage, promote and oversee the individual stock and derivative contract markets. They set and enforce rules regarding trading practices, monitor the financial resources and obligations of participants and supervise the settlement of transactions. There is a system of federal regulatory oversight which requires or prohibits particular rules and practices, approves rule changes, and audits the exchanges' trading and financial surveillance. The Securities and Exchange Commission has responsibility for stocks and options; the Commodity Futures Trading Commission oversees futures.

Margin

Customers of futures commission merchants and broker-dealers in stock markets must post collateral, called "margin," consisting of cash and securities, against their obligations. These obligations are twofold. First, they are loans from a broker-dealer to purchase stock. Second, they are obligations created by a short sale of stock, the purchase or sale of a futures contract and the sale of an options contract. The equity balance of a customer's margin account, equal to the difference between the market value of securities and the amount of the loan or other obligation, is calculated each day. The equity value must be greater than the margin requirement; otherwise the broker-dealer may call for more margin or sell the customer's positions.

The Federal Reserve has final authority for setting initial margin requirements for stocks and options. The individual commodity exchanges have the authority to set margins in the futures contracts traded on their floors.

Clearing

Trades executed on an exchange are guaranteed by a "clearinghouse," whose performance is in turn guaranteed to varying degrees by the clearing members (broker-dealers or futures commission merchants) of that exchange. Most U.S. stock exchanges clear their transactions through a single stock clearinghouse. Similarly, all U.S. options exchanges clear through a single options clearinghouse. In contrast, each of the largest futures exchanges maintains its own clearinghouse.

Trading Strategies

The price of an index futures contract and the price of the stock index portfolio underlying it (often referred to as the "cash index") are directly related. Normally, the price of a futures contract exceeds the price of the underlying portfolio by an amount reflecting the "cost of carry," which relates to the difference between the Treasury bill rate and the dividend yield on the portfolio.

An index arbitrageur attempts to profit when the price difference between futures and stocks is abnormal, either by simultaneously buying futures contracts and selling the index portfolio of stocks or by doing the reverse. When the futures price is at a discount, the arbitrageur engages in index substitution by selling an index portfolio of stocks and replacing it with futures contracts. This is typically done by a pension fund which owns an indexed portfolio of stocks. In executing this arbitrage, the institution takes on whatever greater credit risk there is in owning the futures contract rather than the stocks themselves. When the futures contract is at a premium, the arbitrageur may execute a "synthetic cash" transaction, buying the stock portfolio and selling futures. Typically, a corporation holding short term money market investments would perform this arbitrage to increase its yield.

There are also a number of non-arbitrage trading strategies which involve stocks and futures contracts. First, when trading-oriented investors want to trade on the direction of the market as a whole, they often buy or sell index futures because futures transactions can be executed more quickly and cheaply than transactions involving a diversified portfolio of stocks. Lower transaction costs and lower margin requirements make this possible. Second, larger term investors often find it faster and initially cheaper to initiate portfolio position changes through the futures market. Eventually, the futures position is replaced with stocks. Third, block traders, exchange specialists and investment bankers marketing new stock issues can use index futures to hedge their positions.

Other strategies are designed to react mechanically to market movements by selling in a falling market and buying in a rising market. One such strategy, "portfolio insurance," is designed to allow institutional investors to participate in a rising market yet protect their portfolio as the market falls. Using computer-based models derived from stock options analysis, portfolio insurance vendors compute optimal stock-to-cash ratios at various stock market price levels. But rather than buying and selling stocks as the market moves, most portfolio insurers adjust the stock-to-cash ratio by trading index futures. Indeed, several major portfolio insurance vendors have been authorized to trade only futures and have no access to their clients' stock portfolios. Some option hedging strategies employed by options traders use the same method of buying futures as the market rises and selling futures as the market falls.

■ *This paper was prepared for the Symposium on the Stock Market Crash, held February 8, 1988, at Princeton University and sponsored by the John M. Olin Program for the Study of Economic Organization and Public Policy. We are grateful to all the participants for their comments, and especially to Joseph Stiglitz, Carl Shapiro and Timothy Taylor for their editorial assistance. We would also like to thank all the members of the President's Task Force on Market Mechanisms, much of whose work is reported here, and Bill Fairburn, who provided further research assistance. The interpretations of the Task Force Report contained herein are our own, and do not necessarily reflect the thinking of the other members.*

References

Campbell, John Y., and Robert J. Shiller, "Conintegration and Tests of Present Value Models," *Journal of Political Economy,* 1987, *95,* 1062–1088.

French, Kenneth R., and Richard Roll, "Stock Return Variances: The Arrival of Information and the Reaction of Traders," *Journal of Financial Economics,* 1986, *17,* 5–26.

Greenwald, Bruce C., and Joseph E. Stiglitz, "Externalities in Economies with Imperfect Information and Incomplete Markets," *Quarterly Journal of Economics,* 1986, *101,* 229–264.

Presidential Task Force on Market Mechanisms: Report Submitted January 1988.

Stein, Jeremy C., "Informational Externalities and Welfare-Reducing Speculation," *Journal of Political Economy,* 1987, *95,* 1123–1145.

Summers, Lawrence H., "A Few Good Taxes," *The New Republic,* Nov. 30, 1987, 14–16.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 25–44

Trading Activity and Price Behavior in the Stock and Stock Index Futures Markets in October 1987

James F. Gammill, Jr. and Terry A. Marsh

The events in the equity markets on the days surrounding Monday, October 19, 1987 have been variously described as a meltdown, a crash, a debacle, a break, a collapse, a panic, and "God—or somebody—sending us a signal." In less than a week, equity values fell more than 30 percent; in one day alone, the Dow Jones Industrial Average fell 508 points, or more than 22 percent, while the most active stock index futures contract price fell more than 28 percent. Coincident with the dramatic drop in prices was an equally dramatic rise in trading volume; more than 600 million shares traded on the New York Stock Exchange on the 19th and the 20th, three times the September 1987 daily average. Furthermore, the heavy trading volume occurred amidst numerous opening delays and trading halts for individual stocks; on October 20, the New York Stock Exchange reported 92 opening delays and 175 subsequent trading halts, while the Chicago Mercantile Exchange halted trading in its stock index futures contract for 46 minutes. The demands placed on communications systems by this combination of heavy trading and sporadic halts made it much more difficult for participants to get reliable information about order executions and fund transfers. Growing doubts about the integrity of the trading, clearing, and settlement mechanisms exacerbated the gridlock and threatened the entire system.

These events provoked widespread public concern about the operations of the stock and stock index futures contract markets. Virtually every exchange, industry group, and government agency concerned with the equity markets responded with an investigation of its own. In addition, President Reagan created the Presidential Task Force on Market Mechanisms (hereafter referred to as the "Task Force") to investigate these events and to report to him within sixty days. Our discussion here will rely

■ *James F. Gammill, Jr. is Assistant Professor of Business Administration, Harvard Business School, Boston, Massachusetts. Terry A. Marsh is Associate Professor of Finance, Graduate School of Business Administration, University of California, Berkeley, California. Both authors were staff members with the Presidential Task Force on Market Mechanisms.*

heavily on the report of this Task Force, as well as the reports of the Securities and Exchange Commission and the Commodity Futures Trading Commission. We focus on trading activity on the New York Stock Exchange (NYSE) and on the S&P 500 stock index futures contract traded on the Chicago Mercantile Exchange (CME). Not only is the overall trading activity much greater on these two exchanges than on their rivals, but virtually all index arbitrage activity, which is crucial to the "intermarket" framework presented by the Task Force Report (and described by Greenwald and Stein in the preceding paper), occurs between the NYSE and the CME.

This paper discusses what actually happened during the October market break and the days immediately before. It attempts to lay out a set of stylized facts that describe differing categories of traders and how they behaved and reacted to each other during those days. We believe that this description of what actually happened provides a necessary starting point for financial economists interested in explaining the stock market break.

Background

Over the past twenty years, institutional investors have become increasingly important as owners and traders of stock. During this period, many innovations have been designed to accommodate institutional investors' demands to adjust their exposure to equity risk quickly and efficiently. A common feature of the various reports on the market break is that they focus on these innovations—such as program trading, stock index futures, and portfolio insurance—and their implications for market behavior.

Institutional investors hold diversified portfolios and implementing an investment strategy often requires them to trade many different stocks. In an attempt to reduce transaction costs, institutional investors package their orders to trade many different stocks; another motivation for doing so is that it credibly signals an absence of stock-specific information. The list of orders in the package is called a "program," and program trading is the execution of these orders. The execution of the order for each stock in a program occurs either on the floor of the stock exchange, where a "specialist" is charged with making an orderly market in each stock, or through an "upstairs" block trading desk, where a broker-dealer finds a trading partner or takes the trade himself.

The futures exchanges, which historically developed quite apart from the stock exchanges, are responsible for the innovation of a single market for stock portfolios. Although a stock index futures contract market does not trade the underlying portfolio of stocks, it provides a single market which trades the risks inherent in a standardized stock portfolio.[1] In a futures market, a crowd of competitive marketmakers, called "locals," provide the marketmaking services.

---

[1] In general, a futures contract is an agreement to buy or sell a specified amount of a commodity at a specified price at a specified time in the future. In the case of the S&P 500 futures contract, investors bet on the future value of the S&P 500 index. A futures contract requires two parties—a "long" and a "short," corresponding to a buyer and a seller—who agree on a price. At the expiration of the contract, if the index value is higher than the price specified by the contract, then the long "wins" and the short pays the long the

Stock index futures markets have been a successful innovation. Within three years after its introduction, the Chicago Mercantile Exchange's contract based on the S&P 500 index was the second most active contract market in the entire U.S. futures industry; by 1987, the daily trading volume (measured by the dollar value of the underlying index) was comparable to the daily volume on the New York Stock Exchange. One reason for this success is that the transaction cost of trading a stock index contract is significantly less than the cost of a comparable program trade.

Arbitrage links the stock and futures markets. Index arbitrage involves buying a stock index futures contract and program-selling the underlying stocks (or, alternatively, selling the futures and program-buying the stocks).[2] Unless otherwise impeded, arbitrage activity continues until the relative mispricing between the stock and futures markets disappears. Thus, arbitrage aligns the prices in the futures and stock markets by affecting the balance of buyers and sellers in each market.

In practice, arbitrage opportunities have often appeared, and many firms have tried to profit from them. Some index arbitragers are organized as broker-dealers, in order to operate with high leverage. Others have supplemented more traditional cash management operations within non-financial firms. A third group, often referred to as "substituters" instead of "arbitragers," is composed of institutional investors committed to a strategy of remaining fully invested in an indexed portfolio; these investors earn arbitrage profits by substituting futures for stock, and vice versa, as the opportunities arise.

The development of very liquid markets in stock index futures offers investors the potential for great precision in tailoring the risk profiles of their portfolios. Portfolio insurance is one approach to such risk tailoring; pension and endowment funds have been the largest users of this strategy. In general, portfolio insurance involves reducing equity exposure as stock prices decrease to put a floor under portfolio returns. (For an explanation of portfolio insurance, see the article by Mark Rubinstein in the Fall 1987 issue of this journal.) "Closet portfolio insurers" are investors who follow such a strategy implicitly, perhaps through the use of stop-loss orders. There have been closet portfolio insurers as long as there have been financial markets; however, the new phenomenom in portfolio insurance is the use of highly structured algorithms—dynamic trading strategies—to determine when to buy or sell.

In summary, the relevant experience of the past 20 years is as follows. Institutional investors have become increasingly important in the equity market, and they often have a desire to trade in and out of portfolios. To accommodate them, brokers developed program trading services and the futures industry developed stock index

---

difference between the contract price and the index value; if the index is lower than the contract price, then the short "wins" and the long pays the short the difference between the index value and the contract price.

[2] Arbitrage opportunities depend on the relative prices of the futures contract and the stocks which compose the underlying stock index. The futures price is the value of the underlying stock index upon contract expiration (that is, in the future) at which the buyer and the seller break even. To convert this break-even price to a present value, one must adjust the futures price downwards for time value and for the expected depreciation of the stock index due to dividend payments. This converted futures price can then be directly compared with the current value of the stock index: if the converted futures price is less (greater) than the index value, then the arbitrager buys (sells) the futures contract and program-sells (program-buys) the stocks in the index.

futures markets. These successes, in turn, led to new, structured trading strategies, including portfolio insurance and index arbitrage. These are the players we now examine: the institutions, the arbitragers, the portfolio insurers, and the marketmakers for the stock and futures markets.

The Stylized Facts about Trading Activity

Any meaningful analysis of what happened during the October market break requires some classification of market participants. Table 1 summarizes the trading activity of 11 classes of participants over the period October 15 to October 20. The classification is partly by institutional structure (like pension funds, mutual funds, exchange members, corporations) and partly by trading strategy (like portfolio insurers, index arbitragers, and trading-oriented investors).[3] In sum, the trading activity of the participants listed in Table 1 accounts for roughly one-half of the stock and three-quarters of the futures contracts that changed hands from October 15 to 20. As can be seen, the dollar volume of futures trading is on a par with that in stocks.[4]

Figure 1 and Table 2 provide additional insight into the trading activity in the futures market. Figure 1 shows the daily level of open interest for all S&P 500 futures contracts during 1987.[5] At its lowest level, the open interest is approximately 100,000 contracts, equivalent to $12 billion of stock. From January through September, open interest follows a regular quarterly cycle which conforms to contract maturities: open interest steadily accumulates until the expiration date of the nearest contract, at which

---

[3] The information presented in Table 1 was gathered by the Task Force from four principal sources. First, the SEC asked 28 broker-dealers to provide information about their trading for customer and house accounts. While this information did not identify the customers, it did identify (in many instances) the particular trading strategy which motivated the trade. Second, the Task Force received the exchanges' reconstructions of the transaction sequence. Third, the CFTC maintains a daily surveillance program of "large traders" in each futures market; for example, for the S&P 500 futures contract, the CFTC requires a daily report of every trader with a position of 300 contracts or more. Although these large trader reports are "snapshots," taken at the end of each trading day, they provided the Task Force with considerable information about the trading activities of the major players in the stock index futures markets. Fourth, the Task Force reviewed daily settlement records of the Depository Trust Company (DTC) for information about the daily net changes in the stock positions of institutional traders and broker-dealers.

[4] To clarify the comparison between the NYSE and the CME, the totals shown in Table 1 are in dollars. To estimate the number of shares traded, we suggest the following representative share prices for the four days: $43.31, $42.82, $34.75, and $30.42. These representatives prices are calculated from the daily dollar and share volumes given in Table VI-B-1 of the Task Force Report. The corresponding futures prices are 301.60, 290.00, 234.40, and 215.60; the actual contract volumes are given in Figures 17 and 18 in Study III of the Task Force Report.

[5] To understand the idea of open interest, remember that a futures contract is an unconditional contract; neither the long nor the short can unilaterally back out of it. To be released from a futures contract, one must re-enter the market and make an offsetting trade which "liquidates" the initial (long or short) position. The realized profit (or loss) on the liquidated position is just the difference between the price at which the contract was sold and bought. The "open interest" of a contract market is the number of outstanding contracts which have not been liquidated. Since marketmakers tend to carry very few open positions overnight, the open interest measures the popularity of the contract market among hedgers and speculators. For this reason, it is a better indicator than trading volume of the commercial success of a contract market.

Table 1

Summary of trading activity, October 15 through October 20[a]

	(millions of dollars)			
	Stock (NYSE)		*Futures (CME)*	
	Bought	*Sold*	*Bought*	*Sold*
Oct. 15				
Portfolio insurers	201	257	171	968
Index arbitragers (& substituters)	407	717	717	407
Exchange marketmakers	1,486	1,364	7,530	7,509
Other exchange members	1,849	1,629	*	*
Option marketmakers	*	*	864	998
Trading oriented investors	1,026	1,156	2,236	2,050
Other pension funds	368	190	76	169
Other financial institutions	798	516	57	109
Mutual funds	998	1,419	27	3
Foreign traders	*	*	298	442
Corporate repurchases	*	*	*	*
Total identified above[b]	4,877	4,902	11,976	12,655
Total reported volume	11,400	11,400	18,830	18,830
Oct. 16				
Portfolio insurers	161	566	109	2,123
Index arbitragers (& substituters)	394	1,592	1,705	392
Exchange marketmakers	2,012	1,845	7,125	7,088
Other exchange members	2,124	1,922	*	*
Option marketmakers	*	*	1,254	1,399
Trading oriented investors	1,081	1,446	3,634	3,373
Other pension funds	773	794	294	234
Other financial institutions	1,221	959	126	247
Mutual funds	1,485	1,339	73	11
Foreign traders	*	*	443	479
Corporate repurchases	*	*	*	*
Total identified above[b]	6,733	6,949	14,763	15,347
Total reported volume	14,500	14,500	19,640	19,640
Oct. 19				
Portfolio insurers	449	1,748	113	4,037
Index arbitragers (& substituters)	110	1,774	1,582	129
Exchange marketmakers	3,976	3,239	5,682	5,479
Other exchange members	2,558	2,714	*	*
Option marketmakers	*	*	915	898
Trading oriented investors	1,316	1,751	4,510	2,590
Other pension funds	1,481	875	447	631
Other financial institutions	1,221	1,416	320	525
Mutual funds	1,947	2,168	143	19
Foreign traders	*	*	609	494
Corporate repurchases	411	*	*	*
Total identified above[b]	10,390	11,197	14,320	14,801
Total reported volume	21,000	21,000	18,987	18,987

Table 1: **continued**

	(millions of dollars)			
	Stock (NYSE)		*Futures (CME)*	
	Bought	*Sold*	*Bought*	*Sold*
Oct. 20				
Portfolio insurers	863	698	505	2,818
Index arbitragers (& substituters)	32	128	119	31
Exchange marketmakers	3,240	3,517	2,689	2,718
Other exchange members	1,710	2,002	*	*
Option marketmakers	*	*	544	635
Trading oriented investors	1,495	1,740	4,004	2,765
Other pension funds	920	334	1,070	514
Other financial institutions	2,154	1,579	517	303
Mutual funds	1,858	1,726	51	40
Foreign traders	*	*	418	329
Corporate repurchases	667	*	*	*
Total identified above[b]	11,197	9,594	9,915	10,152
Total reported volume	18,500	18,500	13,641	13,641

*Not available or not applicable

[a] Sources of stock data: Corporate Repurchases, SEC Report, table 6-2; Exchange Marketmakers and Other Exchange Members, Task Force Report, table VI-B-1; all other categories, Task Force Report, figures III-13 and III-14. Sources of futures data: Task Force Report, figures III-15 and III-16.

[b] Due to overlap with other categories, for the stock data the Total Identified Above excludes Index Arbitragers (& Substituters) and Other Exchange Members.

time it falls by roughly 25 percent as most, but not all, expiring positions are rolled over. But in October 1987, the increase in open interest, equivalent to $6 billion of stock, dramatically exceeds that suggested by the historical trends. The sudden upturn in the open interest begins on October 14, peaks on October 20, and returns to approximately its normal cyclical level by mid-November.

The open interest data supports the impression that the users of stock index futures markets placed unprecedented demands for liquidity on the markets during October. Table 2 provides additional information about the source of the demand for new futures positions, short and long; it shows the daily increase or decrease in the aggregate open positions for four groups of traders with large positions.[6] Since the data distinguishes between opening and liquidating transactions, it provides some aggregate data about who is entering and exiting the futures market.

---

[6] The data is from the CFTC Interim Report and is based on the reports of traders with positions of 300 contracts or more. As defined by the CFTC, the four groups are broker-dealers, institutions, other commercials, and non-commercials. The short positions of the institutions reflect portfolio insurance selling. The long positions of the institutions and the broker-dealers reflect index substitution and index arbitrage strategies, respectively.

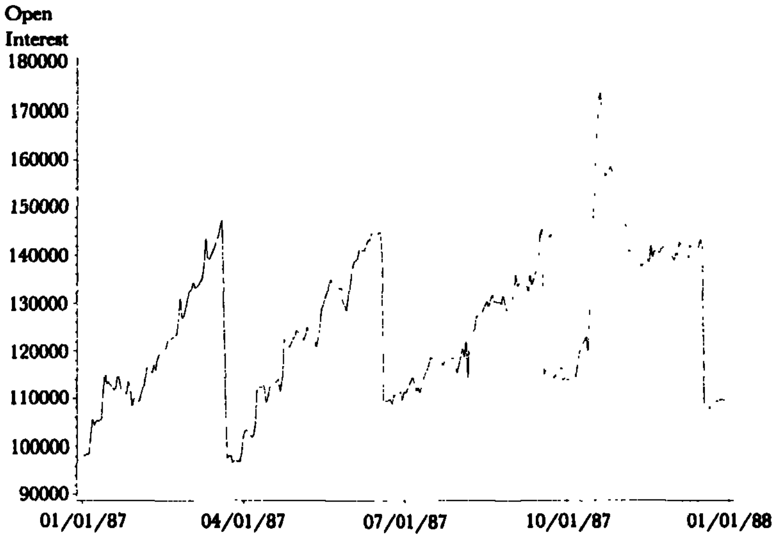


Fig. 1. Daily open interest for all CME S&P 500 index futures contracts.

These tables help to explain why portfolio insurance plays a leading role in most accounts of the market break: of the eleven categories identified in Table 1, the portfolio insurers are the only heavy sellers in both the futures and the stock market. As the market fell between October 14 and October 20, portfolio insurers reduced their net equity exposure in the CME and NYSE markets by roughly $11 billion. On October 19 alone, the gross sales of portfolio insurers on the CME and NYSE were almost $6 billion; on the 20th, gross sales exceeded $3 billion. Table 2 confirms that the selling by portfolio insurers and other institutions extended through Tuesday, the 20th; the net increase in short positions on Tuesday is only slightly less than that of Monday.[7] And as shown by the total net increase of short positions for the two week period in Table 2, selling by institutions, due primarily to explicit and implicit portfolio insurance strategies, significantly exceeded the dramatic increase in open interest.

The other category of trader closely scrutinized by most studies consists of index arbitragers (and substituters), for a similar reason: on October 16 and 19, of the eleven categories in Table 1, index arbitragers were the heaviest net sellers of stocks. Since virtually all arbitrage activity involved buying futures and program-selling stocks, arbitragers were also the largest net buyers in the futures market; thus, arbitragers shifted selling pressure from the futures market to the stock market. But as both the transaction data and open interest data show, the futures buying by the arbitragers was less than the futures selling by the portfolio insurers. Furthermore, the NYSE effectively stopped most arbitrage activity on October 20 by requesting its

---

[7] There is even less difference between the two days since there were three fewer reporting institutions with large short positions for the 20th than for the 19th (42 v. 45). The reporting threshold is 300 contracts.

Table 2:
Analysis of change in open interest ("OI") and large trader positions

Short contract positions			*Increase in short positions reported by category of large trader*			
Date	*Open Interest (contracts)**	*Change in OI*	*Broker-dealers*	*Institutions (Portfolio Insurers)*	*Commercial*	*Non-commercial*
12	123,064	2,336	528	2,343	−2,279	−2
13	119,880	−3,184	230	1,195	−619	−956
14	127,582	7,702	−600	3,139	−1,380	2,548
15	133,696	6,114	683	3,677	2,108	−713
16	146,653	12,957	184	11,590	−2,120	−1,896
19	172,178	25,525	−26	18,220	−1,867	3,704
20	174,184	2,006	−250	17,357	−3,927	−2,563
21	169,934	−4,250	−30	−3,153	356	−1,015
22	158,774	−11,160	41	−3,951	1,254	−506
23	156,650	−2,124	57	−3,712	1,918	−95
26	158,715	2,065	−25	1,015	−389	1,027
Totals:		37,987	792	47,720	−6,945	−467

Long contract positions			*Increase in long positions reported by category of large trader*			
Date	*Open Interest (contracts)**	*Change in OI*	*Broker-dealers*	*Institutions*	*Commercial*	*Non-commercial*
12	123,064	2,336	418	−1,319	−210	73
13	119,880	−3,184	−33	−778	−378	2,922
14	127,582	7,702	2,861	579	1,469	102
15	133,696	6,114	2,001	−248	831	2,879
16	146,653	12,957	1,470	762	−1,462	−1,031
19	172,178	25,525	5,569	4,861	6,644	−109
20	174,184	2,006	8,211	7,507	−3,259	−2,290
21	169,934	−4,250	−2,001	5,113	−2,134	−212
22	158,774	−11,160	1,095	−109	372	−3,075
23	156,650	−2,124	2,916	365	−667	−357
26	158,715	2,065	−2,787	−398	−1,158	−269
Totals:		37,987	19,720	16,335	48	−1,367

*At an index value of 250, 1,000 contracts equals $125 million. From October 12 through October 26, the closing index values ranged from a high of 314.52 on October 13 to a low of 225.06 on October 19.
Source: CFTC Interim Report, tables 17, 29, 40, and 51.

members not to use the SuperDot system to transmit arbitrage program orders electronically to the specialists on the NYSE floor.[8]

While the portfolio insurers and index arbitragers have received the most attention, the behavior of three related categories—the exchange marketmakers, the other exchange members, and the option marketmakers—is also noteworthy. The overall impression is that their trading is consistent with their marketmaker designation. All three operate as very short term intermediaries, bridging the imbalances in the incoming order flow. For the most part, they were active traders within the day and their trading was roughly balanced between buying and selling. On the 19th, most were net buyers; on the 20th, most were net sellers.

---

[8] SuperDot is the electronic system by which market and limit orders for stocks are routed to specialist posts or member booths on the NYSE trading floor, and confirmations of executions are returned. The Opening Automated Report Service (OARS) of SuperDot accepts pre-opening orders and automatically pairs buys and sells for an auction-like determination of opening prices.

More specifically, the exchange marketmakers are the specialists on the NYSE floor and the locals on the CME floor. The aggregate behavior of the marketmakers on the two exchanges was similar: on the 19th, the specialists were net buyers of $500 million while the locals were net buyers of $200 million; on the 20th, both the specialists and the locals were net sellers of $300 million and $100 million, respectively. The similarity extends to gross trading volume as well; over the 19th and 20th, the gross purchases of the NYSE and CME marketmakers were $7.7 billion and $8.7 billion, respectively.

The category of other exchange members includes all trading for exchange members' accounts which is not done by specialists and locals for their own accounts. This category includes two types of firms: those firms that provide customers with marketmaking services through their block trading desks away from the exchange trading floor (that is, "upstairs"), as well as those member firms that trade primarily for their own accounts. The influence of the second type of firm is evident in the fact that in aggregate the "other members" of the NYSE were net sellers on both the 19th and 20th, while a separate SEC survey of fourteen large block traders found them to be net buyers on the 19th, purchasing 33.3 million shares and selling 27.6 million shares (approximately $1.2 billion and $1.0 billion, respectively).

Although the trading of option contracts is not the focus of this paper, some marketmakers at the option exchanges used the stock index futures market to hedge their positions.[9] The greatest activity occurred on October 16 when many option contracts expired, requiring the marketmakers to liquidate their hedge positions in the futures market.[10] On all days, the buys and sells of the option marketmakers in the futures market were relatively balanced.

The category identified as trading-oriented investors is the second most active category, after the exchange marketmakers. The category includes broker-dealers, pension funds, banks, and other institutional investors with very short investment horizons; they are aggressive traders who may open and close positions repeatedly during a single trading day. Their behavior during these four days resembles that of the index arbitragers in that they were net sellers of stock and net buyers of futures; both groups responded similarly to the relatively cheap prices in the futures market vis-à-vis the stock market. During the week of October 12, much of their trading relates to the expiration (on October 16) of index option and futures contracts. As one might expect of those with short-term horizons, their heavy trading did not lead to dramatic changes in their net holding of equities and futures.

---

[9] An option contract is similar to a futures contract, in that both are agreements to buy or sell a specified amount at a specified price at a specified time in the future. But an option contract differs from a futures contract in that one party to an option contract—either the long or the short—unilaterally decides whether or not the transaction described by the contract will occur. If the exercise right, or option, is held by the long, it is a "call option," and if the option is held by the short, it is a "put option." The party that grants the exercise right is called the option "writer." In return for granting the option, the writer receives a payment, or "premium," from the option buyer when the option contract is initially established.

[10] Upon expiration of a stock index futures or option contract, all contract positions are liquidated based on the index value calculated with that day's closing stock prices. This "cash settlement" feature has two effects: first, it increases the importance of the closing stock prices on the expiration day, and second, it facilitates short term speculation on relative and on absolute stock index movements. Friday, October 16, was an expiration day for several broad-based stock index option and futures contracts (not including the CME's S&P 500 futures contract).

As their names suggest, the categories of other pension funds and other financial institutions are somewhat heterogeneous and thus it is more difficult to clearly identify the nature of, or motive for, their trading. The category of other pension funds is composed of those funds not following explicit portfolio insurance strategies; consequently it includes "closet" portfolio insurance selling (in the futures market), index substitution (that is, buying futures and selling stock), and "value buying" (that is, buying stock and futures after their prices drop sharply).

The category of other financial institutions—which include banks, insurance companies, and other institutional investors—generally behaved like the other pension funds in that they were net buyers of stocks and, in particular, heavy buyers on October 20 when most of the decline in stock prices had already occurred. Their participation in the futures market was minor; they were net sellers for the first few days and then buyers of futures on October 20.

Two other categories which employ a wide range of trading strategies are mutual funds and foreign traders. Mutual funds, as a group, were buyers and sellers of comparable amounts of stock during each day. However, this appearance of overall balance conceals large differences in the net buying and selling of individual mutual fund firms. In general, the mutual funds that were the heavy sellers were those that made it easiest for their shareholders to switch out of equity funds, through such innovations as unlimited telephone transfers and extended hours for customer service; these funds sold in response to realized and anticipated demand for redemptions of equity fund shares. In contrast, the funds that were net buyers operated in a more traditional manner, and viewed the week of the 19th as an opportunity to buy stocks at favorable prices. As Table 1 shows, mutual funds were not strong participants in the futures markets, largely because many funds are not authorized to trade futures contracts.

Foreign traders were relatively heavy players in the futures market, but their net position changed little over these four days. They were net sellers of futures on October 15 and 16, and net buyers of futures on the 19th and 20th.

In contrast to the preceding ten categories, the final category, corporate repurchases, consists exclusively of buyers. At midday on the 20th, when dark rumors about the exchanges and their member firms were rampant, the well-publicized announcements of repurchases appeared to provide much-needed support for stock prices in general. The salutory effect of these repurchases and announcements on a market in disarray is reminiscent of a similar effect brought about by well-publicized purchases, on the afternoon of October 24, 1929, of U.S. Steel and other blue-chip stocks by a pool organized by the leading Wall Street banks. Also of note is the difference between the announced size of the repurchase programs and the actual shares bought. Of the S&P 500 corporations, 129 repurchased 21.94 million shares of stock on October 20, and 90.39 million shares (for over $2 billion) during the week. The repurchases on the 20th were about one-tenth of the total repurchases announced that day, which suggests that if allegations that the repurchase programs were implemented under duress are true, those doing the arm-twisting must have been satisfied with the announcement effects.

Within the eleven categories described here, then, the trading patterns around "Black Monday" might be summarized as follows. Portfolio insurers were heavy sellers of futures and stocks. Index arbitragers and their institutional counterparts, index substituters, were heavy buyers of futures and sellers of stock. A handful of firms and institutional investors were active intraday traders, opening and closing positions around major price movements. Marketmakers bought on the 19th when confronted with overwhelmingly large imbalances of sell orders, but in aggregate they were net sellers on the 20th. Other institutional investors were strong net buyers, particularly once prices declined considerably. Corporations provided much needed support after the 19th, through well-publicized announcements of share repurchase programs. Other actors played the market, but did not change their net position dramatically.

We close this section with three additional observations and interpretations of the stylized facts. First, the amount of trading done by the most active participants is worth mentioning. On October 19, when roughly $21 billion of NYSE stocks and $20 billion of S&P 500 futures contracts were traded, four investors accounted for more than $4.2 billion of sales: one portfolio insurer sold $1.1 billion of stocks and $600 million of futures; two other portfolio insurers sold a total of $1.7 billion of futures; and one mutual fund sold more than $800 million of stock near the Monday open. In short, this handful of traders formed a significant part of the selling pressure on that day.

Second, the lack of heavy net selling by option marketmakers is intriguing. If options were written by marketmakers who dynamically hedged their positions in the futures and stock markets, then we would have seen much greater selling by them.[11] Several possible explanations for the data are: selling by option marketmakers went relatively undetected by the studies; option marketmakers were hedged through the purchase of other options; and option marketmakers were unhedged and suffered substantial losses as a result. We know that the last explanation is true to some extent;[12] we suspect that there is some truth to the first two explanations as well.

---

[11] The writer of a call (put) option can hedge an investment position by establishing a long (short) futures position that has the same immediate exposure to stock index price movements as the option contract itself. The appropriate size of the futures contract position, referred to as the "hedge ratio," changes with time and with the value of the underlying index. In particular, as the index falls in value, a call option writer must reduce his long position and a put option writer must increase his short position in order to remain properly hedged. Thus, a fall in the value of the index triggers futures market selling by call and put option writers who follow dynamic hedging strategies.

[12] The SEC reports that First Options, the clearing firm for the largest number of stock index option marketmakers, established in October 1987 a reserve for bad debts associated with the marketmaker deficits of approximately $91.6 million, most of which is attributable to nine marketmakers.

Our third additional observation is that there is an important difference in the nature of the trading activity in the CME stock index futures market between Monday, October 19, and Tuesday, October 20. Returning to Table 2, note that the reporting institutional investors increased their short positions by 17,357 contracts on the 20th, comparable to their increase on the 19th of 18,220 short positions. Yet the total increase in open interest on the 20th was only 2,006 contracts, much less than the previous day's increase of 25,525 contracts. Thus, on net, the existing shorts not identified as reporting institutional investors (essentially, the existing shorts who were not portfolio insurers) liquidated 15,351 contracts, equivalent to roughly $2 billion of stock.

Why is it that portfolio insurers continued to increase their short positions on October 20th, while most other investors liquidated their positions? We can suggest four different interpretations, all of which could be at least partly accurate. The first possibility is that at least part of the non-institutional selling to establish new short positions from October 14 through October 19 was by those who (correctly) anticipated subsequent selling by portfolio insurers, and that their "anticipations" were realized on October 20. The second is that the inability of index arbitragers and substituters to use the SuperDot system to implement their stock trades on October 20 kept them from buying futures and thus eliminated an important source of new buying. The third explanation is that positions were liquidated and new positions discouraged because of concerns about abrupt, unscheduled trading halts which would freeze the traders' positions. A fourth possible explanation, similar to the third, is that heightened concern and uncertainty about the integrity of the CME clearinghouse encouraged many holders of contract positions to liquidate and thus reduce their exposure to credit risk.[13]

Each explanation suggests a different problem and remedy. The first explanation implies that, at least for this realization, there was a cost to portfolio insurance strategies which resulted from aggressive traders selling ahead of anticipated portfolio insurance sales. Further, insofar as this explanation is predicated upon the predictability of portfolio insurance sales, such a cost is not likely to be avoided by "sunshine trading"[14] and the like. The second explanation suggests that the CME's stock index futures market is directly affected by the NYSE's actions with respect to their members' access to the SuperDot system. The third explanation implies that the feedback effects of unscheduled and, by extension, scheduled "circuit breakers" need to be considered carefully; both illustrate the need for greater cooperation and coordination between the exchanges of their policies and practices regarding inter-market trading. The fourth explanation emphasizes the importance of maintaining the participants' utmost confidence in the integrity of the clearing and settlement systems.

---

[13] Before accepting either of the last two explanations, one needs a story as to why concerns about trading halts or credit risks did not deter the portfolio insurers from selling. Our conjecture is that the portfolio insurers recognized these concerns but felt bound by their contractual obligations.

[14] "Sunshine trading" refers to the announcement that a trade of a given size will occur at a specified time in the future. Sunshine trading was employed by Leland, O'Brien, Rubinstein (a portfolio insurance vendor), but is currently prohibited because it is believed conducive to off-floor trading.

Intraday Trading Activity and Price Behavior

A common feature of the reports of the Task Force, the SEC, and the CFTC is detailed attention to the chronological sequence of index-related trading activity on the NYSE and CME. This approach follows earlier analyses of similar events; the October 1987 market break was, in fact, the third such price decline within a 14 month period to receive special attention from the SEC and the CFTC.[15]

The chronologies presented by all three reports feature calculations by half-hour bracket of the concentration of orders motivated by a similar strategy, and they compare these concentrations with the price declines during the half-hour brackets.[16] One objective of these comparisons is to examine a commonly offered explanation for the market break: Selling by portfolio insurers to hedge existing stock positions depresses futures prices, which in turn attracts index arbitragers who buy the cheaper futures and sell stocks, thus transferring the selling pressure to the stock market. This literal explanation is tested by examining whether price declines in the futures market are correlated with portfolio insurance selling, and whether price declines in the stock market are correlated with index arbitrage and substitution selling.

As each of the three reports acknowledge (to varying degrees), the correlation between index arbitrage selling and stock price declines is considerably stronger than the correlation between portfolio insurance selling and futures price declines. In fact, portfolio insurers appeared to sell as much or more when the futures price was firm or rising than when it was dropping rapidly. Figure 2 shows the half-hourly S&P 500 price changes and index arbitrage net selling for October 19, while Figure 3 shows half-hourly portfolio insurance net selling and S&P index futures price changes on the CME. The inverse association between index arbitrage selling, which is positively correlated with the difference between the prices of stock and index futures, and stock price changes simply reflects index arbitrage in operation as it transfers options and futures market short selling to the stock market. The positive association between net futures (or stock) sales by portfolio insurers and price changes does not, at face value, confirm the popular opinion that portfolio insurers "broke" the market.

However, conclusions regarding the market break based on an analysis of the contemporaneous association between trading activity and price changes can be misleading. One obvious reason for this is possible inaccuracies in the data. The times

---

[15] The two other breaks occurred on September 11–12, 1986, and January 23, 1987. See "The Role of Index-Related Trading in the Market Decline on September 11 and 12, 1986," by the Division of Market Regulation, U.S. Securities and Exchange Commission (1987). There are some obvious differences between these and the October 19–20 breaks. For example, a price run-up (52.60 points) on January 22 preceded the January 23 break, which was in turn accompanied by a general lack of portfolio insurance selling.

[16] For the October 1987 break, the SEC and CFTC jointly requested data of eleven broker-dealers and four portfolio insurance vendors who executed index arbitrage and substitution orders, program trades, and portfolio insurance strategies for their customers and their house accounts. Appendix B of the SEC Study lists each program trade identified by the broker-dealers; it includes the time the order was entered, the size of the order, the underlying strategy, and whether it was a customer or proprietary trade. Appendix C of the SEC Study lists the time and size of portfolio insurance orders submitted to the CME S&P stock index futures market.

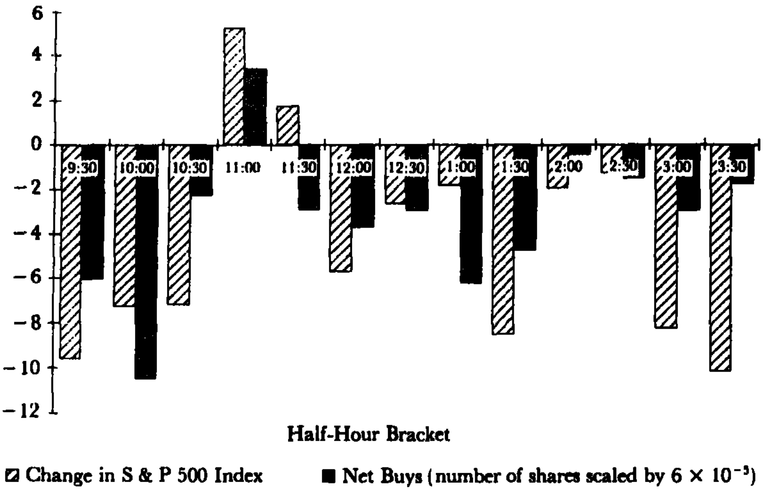


Fig. 2. NYSE—October 19 index arbitrage net buys and S&P price changes.

reported by the broker-dealers are order entry times, as opposed to order execution times. During October 19 and 20, equipment constraints—most notably, of printers, disk storage, and software unable to deal with error messages—delayed the transmission of many orders to the specialists' posts, particularly those not equipped with DisplayBook. The Task Force's examination of the SuperDot files yielded evidence not inconsistent with the GAO's report that 10 minute to 75 minute delays in order execution were experienced in the afternoon of October 19 and on October 20. Thus, plausible adjustments to the order entry times employed by the SEC and Task Force

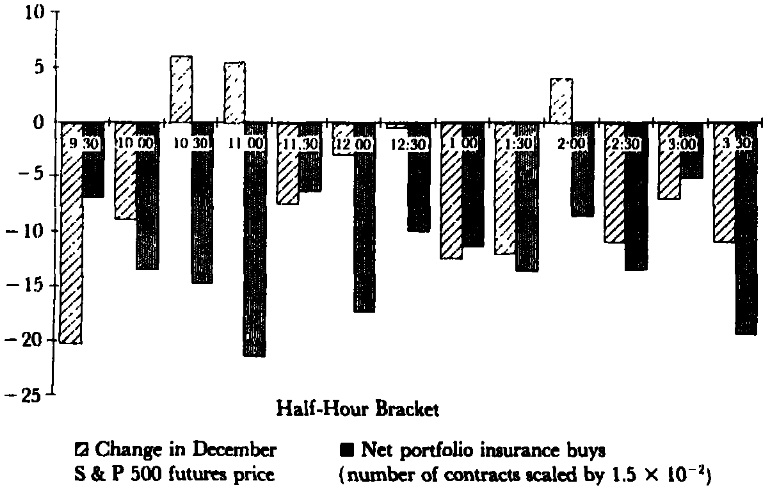


Fig. 3. CME—October 19 net portfolio insurance buys and price changes.

could easily lead to a stronger negative association between trading activities and price changes than those suggested by Figures 2 and 3.

A more significant problem with the "popular" explanation described above is the presumption that price declines should necessarily coincide with portfolio insurance and index arbitrage selling. If the market mechanism extracts a cost of immediate order execution—that is, if the demand or supply curve of futures contracts or stock is not perfectly elastic over short intervals—then the sequence of trades over that interval matters. A trader who correctly anticipates imminent order flow into the market can realize nearly immediate trading profits by opening a position ahead of the incoming orders and then liquidating the position by trading with incoming orders.[17]

If such anticipatory trading precedes selling by portfolio insurers, and if the market mechanism exhibits "price pressure" as a cost of immediate order execution, then the selling which opens the anticipatory position will coincide with the drop in price. Depending on the trader's eagerness to unwind his position, the subsequent trade with the portfolio insurer need not result in any further price movement in order to realize a profit for the anticipatory trader.

The Task Force Report identifies anticipatory trading as a factor in the market break. Portfolio insurance strategies were well understood by professional market participants; a reasonable assumption was that portfolio insurers would want to reduce equity exposure by 2 percent of protected assets for each 1 percent decline in the stock market. Thus, assuming $60 billion under protection by portfolio insurance, the 10 percent decline in the market between October 14 and October 16 corresponded to a desire to reduce equity exposure by $12 billion, increases in transactions costs aside. This $12 billion corresponded to approximately 300 million shares of stock or 100,000 S&P 500 futures contracts, and would have been higher if some insurers increased their hedge ratios to account for higher volatility, or if they were insuring assets net of liabilities (typically pension obligations) which increased in value as interest rates declined. But through October 16, portfolio insurers had sold only $4 billion, and there was considerable awareness of the "overhang" of incomplete portfolio insurance selling. This overhang supposedly encouraged additional selling on the morning of the 19th. The open interest data in Table 2 also suggests that portfolio insurers sold, to some extent, to traders who had previously established short positions with the intention of closing them out quickly.

To understand fully the relationship between price changes and trading activity, one must know the identities of both sides of a trade as well as their overall trading strategies. In the zero sum trading game which characterizes trading in secondary markets, "selling pressure" (or "buying pressure") is a relative, not an absolute, concept and it must be interpreted in context of the other side of the trade over a given interval. Current transaction records do not permit us to determine which side of a trade "initiated" it, nor who was on the contra side. And of course, even if they did, singling out one side of a trade as the initiating one would be difficult in a rational expectations equilibrium unless shocks to expectations could somehow be observed and trading strategies derived.

---

[17] We recognize that there is a gray area between legitimate speculation on short term price movements and illegitimate front-running whereby brokers, qua principal, trade ahead of their customers' orders in violation of some well defined principal-agent contract. See the CFTC Report (1988, pp. 197-200) for a discussion and further references.

The behavior of specialist positions suggests whether there is "selling pressure" on prices: given an assumption about a specialist's optimal inventory position, models of marketmaking such as those of Garman (1976) and Ho and Stoll (1981) predict movements in bid and ask prices (and therefore "price pressure") in response to trades in which a specialist participates. The marketmaking models are inherently stock-specific, so a multivariate analysis of price changes and specialists' position movements would provide an opportunity to check whether adverse selection was at all important on October 19 and 20 and, perhaps most important, it would enable us to incorporate cross-sectional information regarding the generation of order flow (for both market and limit orders).

Was there "pressure" on the NYSE specialists and the CME locals? The Task Force Report includes a univariate analysis of half-hourly trading activity of the specialists in 31 NYSE stocks and of the locals (in aggregate) in the S&P 500 futures contract. Net buying by these 31 specialists on October 19 and 20 is negatively correlated with price changes, as is CME local buying. On October 20, for example, specialists were heavy net sellers in the early morning and after 1:30 p.m. when prices were rising, and net buyers as prices declined in the late morning; the transactions of the CME locals followed the same pattern, though they were net buyers in mid-afternoon (for further discussion, see the Task Force Report, Study VI). The net buying occurred simultaneously with a significant increase in quoted bid-ask spreads. On October 19 and 20, 30 percent of quoted spreads were a quarter or less, for example, compared with 71.7 percent in September 1987 (Task Force Report, Study VI, Tables B-3 and B-6). Instead of using quotes which were potentially unreliable on October 19 and 20, we also approximated them by examining marketmaker trading profits in our analysis. Marketmaker buying when prices decrease and selling when prices increase (in the interests of stabilizing the market, of course) could, particularly with the wider bid-ask spread, generate a considerable trading profit to offset, or more than offset, losses on initial inventory position.

Of course, specialists typically only participate in about one quarter of all NYSE trades. Also, as noted earlier, off-floor members, who accounted for 16 percent of all purchases and sales in 1986, were in aggregate net sellers of stocks on October 19 and 20. The limitation in examining trading activity and price behavior was in determining who the other traders were, even for the Task Force which had full access to current audit trail and trade reconstruction technologies. Available data does show that an abnormally large number of limit buy orders cumulated in SuperDot were executed on October 19 (though the GAO Report states that the As-Of-Status-System which reports the executions was down on October 20). The change in the positions of the large traders analyzed earlier also offers some limited additional information on trading activity.

Another obstacle to associating trading volume and price changes on October 19 and October 20 is that a lot of uncertainty existed about the market mechanism itself.

Such uncertainty was presumably imputed to trading costs and "effective prices" (both for limit and market orders) by investors, but we cannot observe this directly. Each of the reports discussed at least some indirect evidence of market breakdown on the 19th and particularly the 20th—days on which, as we've discussed above, activity was unprecedented. The evidence included rumored and actual exchange closings, trading halts, delays in execution, less reliable quotes, and greater perceived risks in the clearing process. Perhaps the best evidence of the effects of these de facto "circuit breakers" is the substantial discount between the S&P 500 futures and the index level (after correcting for non-trading effects).[18]

Conclusions and Implications for Financial Economics Research

The Task Force mandate was to study whether and how the interaction of new investment technology and the market mechanism contributed to the market's volatility in October. This focus seems particularly justified given the sharpness of the price break in the week of October 19 and reasonably widespread suspicions that portfolio insurance and/or index arbitrage contributed to volatile swings in prices back in September 1986 and January 1987.[19]

Whether the price movements in October 1987 reflect only problems of short-term illiquidity and market disarray or, given that prices have not completely rebounded, are partly symptomatic of some fundamental change in the market (for whatever reason) cannot be determined from a month's worth of data on essentially one realization.[20] The events described in the Task Force Report appear to have more import for transactions-level models of rational learning and, more particularly, the way in which "rational" panic can occur when markets and their feedback mechanisms break down. The Task Force Report provides statistics for the study of models in which it is rational for some subset of investors, be they portfolio insurers or "closet" portfolio insurers—who can be observationally equivalent to "panicked" investors—to sell when prices fall.

Certain aspects of the events of October 1987, and of the triple-witching days in September 1986 and January 1987, do seem everyday enough to warrant new emphasis in research regarding the microstructure of financial markets itself. At the heart of an examination of the sharp market shifts is a perspective about the size of the markets. From one perspective—the market value of its listings—the NYSE is a more than $2 trillion market, yet the total volume of trade on the three heaviest days combined amounted to less than $60 billion, or 3 percent of the outstanding equity. Even if the additional sales (and thus purchases) of stock that might have been discouraged by high transactions costs on October 19 and 20 were counted, the fact that trading "only" a small percentage of the outstanding stock can move prices more than 20 percent indicates to some that the market mechanisms, and only the market mechanisms, are seriously deficient.

---

[18] Of course, it might appear that some firms could still have arbitraged the discount. However, the few firms that were set up to do index arbitrage manually were restricted to some extent by the uptick rule. (Evidence that the uptick rule was a binding constraint consists of the number of EFP's (Exchange-For-Physicals) that took place.) Many of the funds which could do index substitution had already swapped, and mutual funds were constrained by the "short-short" rule.

[19] These earlier price swings are important because their treatment in computation of volatilities largely determines whether one concludes that volatilities have increased.

[20] Unfortunately, evidence subsequent to the event probably won't throw much light on longer-run price elasticity of demand because "all else" has hardly remained equal. Aside from revisions in priors occasioned by the price movements themselves, there is evidence of increased uncertainty concerning corporate cash flows in the week ending October 16: higher implied equity volatilities and interest rates, proposed takeover legislation, financial press reports of the concerns of institutional investors, and so on. Depending upon how long increased uncertainty takes to revert to its long-run steady-state and how it impacts equilibrium risk premiums, it is capable of both acting as a catalyst to the events of the week of October 19 (for example, Gennotte and Marsh, 1987), and explaining "new" stock price levels.

From the perspective of the marketmaking community, however, the activity of October 1987 looks much different. The amount of equity risk which potentially could be transferred quickly and easily is much less than the outstanding stock, for reasons analogous to why it would be unrealistic to assume that a large portion of the housing stock could turn over on short notice, even if that stock were homogeneous. In this light, the events of October indicate considerable problems in readily attracting orders from those who are willing to adjust their equity positions in response to a significant change in price and that the market mechanisms need to be improved accordingly. There remains the problem of representing or discovering the demands of marginal investors who are willing to trade in response to the demands for immediacy which innovative trading strategies, such as portfolio insurance, may generate. In the same vein is research such as Grossman's (1988) which claims that the problem in lining up the contra side of the trade is that synthetic portfolio insurance strategies do not reveal the true extent of the underlying demand for hedging.[21]

The first perspective assumes perfectly elastic demand curves for equity and considers the entire equity stock to be available for trading; the latter regards trading as a specialized financial service which is accommodated primarily by professional marketmakers working with a limited capital base. In the latter case, "price pressure" is a genuine phenomenom, since it is the mechanism by which the professional marketmakers manage their inventories and realize trading profits. There are several potential ways of enhancing marketmaker capacity by allowing investors to better represent their willingness to trade, such as introduction of "participating orders" along the lines suggested by Black (1971), and other changes discussed in the preceding paper by Greenwald and Stein.

---

[21] Piecemeal evidence from October 1987 does not suggest that unpredictable hedging demands were a problem. In fact, the indictment of portfolio insurance was that, if anything, it generated a large, predictable "overhang of selling pressure...to crush the equity markets in the following week." As we discussed earlier, if there was a surprise, it was probably how low portfolio insurance sales were. Also, maximum selling demands were again claimed to be predictable on October 20, but stocks opened up (or didn't open). From an empirical viewpoint, the contemporaneous half-hourly association between portfolio insurance sales (order submissions) and prices was positive, not negative, on October 19 and 20. Many have concluded that portfolio insurance turned out to be unexpectedly expensive, but so may have been the case for put option writers if the insurance had been implemented via put options instead of dynamic trading strategies.

It is obvious that the emphasis on the bridging function of marketmakers can be reduced if market mechanisms are redesigned to enhance the discovery of the ultimate contra side of trades. The conceptual economic questions concerning the best design for achieving this are among the most interesting ones highlighted by October's experience, and the subject of the Task Force's most controversial recommendation concerning circuit breakers. The rationale behind well-designed circuit breakers, properly coordinated across markets, is that they provide time to find the contra side of trades, at a better expected execution cost, for those investors willing to wait. Blocks of 10,000 shares or more, positioned "upstairs" by NYSE member firms and then crossed on the floor, already account for more than 50 percent of NYSE volume. Price limits are one obvious circuit breaker, but have well-known disadvantages such as feedback to trading behavior, particularly as the limits are approached, and the locking-in of positions when prices go up or down the limit.

Discrete-time auctions with pre-announced rules along with continuous trading, or in place of continuous trading when imbalances cause it to halt, seem to be an interesting alternative which already exists to varying degrees in Japan and other countries, in the U.S. government bond market, and even in the design of the OARS system used on the NYSE for opening stocks (see Wunsch (1987) for discussion). Their design poses several challenging research questions. For example, since markets are usually deepest when trading is concentrated, what will be the liquidity characteristics of the auction and continuous trading markets if they exist simultaneously?

A second topic is greater consideration of the strategic effects of a market based on "price pressure" trading mechanisms. In such markets, the sequence of trading matters and there is a strong economic incentive to anticipate the liquidity demands of others. The conditions under which deliberate price manipulation by uninformed traders (an important issue if numerous anecdotal accounts were to hold up) will be successful would follow from such consideration of strategic effects.[22] Similarly, front-running, in which principals trade ahead of customer orders, presumably because of an imperfect (though perhaps currently second-best) principal-agent contract, becomes more important when trades motivated by liquidity considerations influence prices.[23]

Understanding the transaction pricing process requires a specification of the motivations and strategies of both sides of the trades. Such a framework is necessary to assess the limitations of trading and hedging models which assume an exogeneously determined price dynamics, and to develop further models of equilibrium in which investors account (or, in irrational models, don't account) for the impact that their trading strategies have on prices. For example, as the total amount of insured funds decreases, the cost of insurance is reduced for the remainder—the characterization of this process as "self-correcting" by the Task Force Report and the financial press presumes that the reason that funds under insurance have decreased is that costs were previously underestimated.

---

[22] One can easily construct models of price manipulation that involve cornering in futures markets, influencing new issue prices relative to those guaranteed, influencing stock prices to generate higher payoffs on derivative instruments, and so on, but it is more difficult to understand manipulation of an asset's price by trading only in that asset with no price-contingent contracts.

[23] Of course, front-running of principal buy and sell recommendations for stocks is sometimes alleged, and investment firms point to "Chinese walls" between security analysis sections and upstairs trading desks as preventatives.

Of course, this identifies only a few of the research topics which the events of October 1987 will stimulate. We are excited about the prospects for research initiatives based on the data that the market break left in its wake.

■ *The first draft of this paper was prepared for the Symposium on the Stock Market Crash, held February 8, 1988, at Princeton University and sponsored by the John M. Olin Program for the Study of Economic Organization and Public Policy. We thank the staff members of the Presidential Task Force on Market Mechanisms, the Securities and Exchange Commission, and the Commodity Futures Trading Commission who contributed to the Reports cited in this paper. The interpretations of the data are our own, and do not necessarily reflect the thinking of the Task Force.*

References

Black, Fischer, "Toward a Fully Automated Exchange, I," *Financial Analysts Journal*, July-August 1971, 1-6.

Garman, Mark B., 1976, "Market Microstructure," *Journal of Financial Economics*, 1976, *3*, 257-275.

Gennotte, Gerard, and Terry A. Marsh, "Variations in Economic Uncertainty and Risk Premiums on Capital Assets," unpublished manuscript, U.C. Berkeley, April 1987.

Grossman, Sanford, "An Analysis of the Implications for Stock and Futures Price Variability of Program Trading and Dynamic Hedging Strategies," paper presented to the Berkeley Program in Finance Seminar, Santa Barbara, March 1988.

Ho, Thomas, and Hans R. Stoll, "Optimal Dealer Pricing under Transactions and Return Uncertainty," *Journal of Financial Economics*, 1981, *9*, 47-73.

The Presidential Task Force on Market Mechanisms. Submitted to The President of the United States, The Secretary of the Treasury and The Chairman of the Federal Reserve Board. Washington, D.C.: USGPO, January 1988.

Rubinstein, Mark, "Derivative Assets Analysis," *Journal of Economic Perspectives*, Fall 1987, *1*:*2*, 73-93.

U.S. Commodity Futures Trading Commission, "Final Report on Stock Index Futures and Cash Market Activity During October 1987." Division of Economic Analysis, The Division of Trading and Markets. Washington, D.C.: USGPO, January 1988.

U.S. Commodity Futures Trading Commission, "Interim Report on Stock Index Futures and Cash Market Activity During October 1987. Division of Economic Analysis, Division of Trading and Markets. Washington, D.C.: USGPO, November 9, 1987.

U.S. Securities and Exchange Commission, "The October 1987 Market Break: A Report by the Division of Market Regulation." Washington, D.C.: USGPO, February 1988.

U.S. Securities and Exchange Commission, "The Role of Index-Related Trading in the Market Decline on September 11 and 12, 1986." Division of Market Regulation. Washington, D.C.: USGPO, March 1987.

Wunsch, R. Steven, "Weekly Commentary: Stock Index Futures," Kidder, Peabody & Co., Financial Futures Department, 1987.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 45–50

Comments on the Market Crash: Six Months After

Hayne Leland and Mark Rubinstein

Six months after the market crash of October 1987, we are still sifting through the debris searching for its cause. The most likely antecedent, very significant new news, is difficult to find. Most of the negative news often cited, such as the rising federal budget and balance of payments deficits, should have been well digested by the market before October 19. With hindsight, it now appears that, unlike the October 1929 crash, the 1987 crash did not presage a general economic decline.

However, one piece of news, the prior behavior of the market itself, was new. In the second week before the crash, the S&P 500 index had fallen 5.2 percent. This decline accelerated in the week preceding the crash when the index fell 9.2 percent culminating with one of the largest one-day declines on record (and the largest in terms of Dow Jones Industrial Average points) of 5.2 percent on Friday, October 16. The market had suddenly exhibited a jump in volatility. This was mirrored in the volatilities implied by the market prices of S&P 500 index options, which increased from 22 percent per annum on October 15 to 30 percent at the close on October 16. Over the weekend, many investors were absorbing the fact that market volatility had at least temporarily increased. Under these conditions, the nervousness of some small investors surfaced in the form of mutual fund redemptions. If you had called one large mutual fund, which offers a telephone transfer service, you would have found the telephones busy at 11:00 p.m. Saturday night. With foreign markets, which begin trading prior to the NYSE, opening down early Monday morning (−2.5 percent in Japan and −10 percent in London), the stage was set for something dramatic.

Unprecedented order imbalances in several large stocks delayed openings on Monday in some cases for as much as two hours. The S&P 500 December futures contract, a much more sensitive barometer of the market under these conditions than

■ *Hayne Leland and Mark Rubinstein are Professors of Finance, Graduate School of Business, University of California, Berkeley, California. They are also consultants to the investment management firm of Leland O'Brien Rubinstein Associates.*

the S&P 500 index itself, gapped open down 6.5 percent. By that measure the market had already declined by more than on any date since the 1929 crash and the day was just starting. As the 19th wore on, investors witnessed symptoms of market failure and were frightened by rumors that the NYSE would close. They also worried that other investors had come to believe the market was overvalued. Fear fed upon fear as investors en masse rushed to sell their stocks. By the end of the day, the S&P futures contract closed down 29 percent and the S&P index closed down 20 percent on NYSE volume of 604 million shares valued at over $20 billion. This is a story that belongs as a new chapter in Charles Mackay's *Extraordinary Popular Delusions and the Madness of Crowds*.

But is that what happened? Perhaps, as the Brady Commission report argues, what would have been a relatively minor decline was turned into a rout by the orders of just a few large traders, many of whom were portfolio insurers. These traders use computerized strategies which dictate buying after the market rises and selling after the market declines. The extreme version of this argument maintains that portfolio insurance first pushed the market up because the "insured" investors were willing to buy more stock since they had planned for systematic sales in market declines.[1] In other words, they could afford to take greater risks in rising markets because portfolio insurance offered a disciplined way of avoiding risk in declines. According to this argument, this additional demand in the year prior to the crash accounted for a good part of the 1000 point rise in the Dow Jones Industrial Average.[2]

Then, as the market went into decline, the insurers began to unwind their positions forcing the market in reverse. But, this time the amount of capital devoted to portfolio insurance strategies had reached a critical mass of about $60 to $80 billion, enough for portfolio insurance sales to fuel themselves. With the sudden fall in the market during the last half hour of trading on October 16, many insurers found themselves with an overhang of unfilled sell orders going into Monday. In addition, several smart institutional traders knew about this overhang and tried to exit the market early Monday before the insurers could complete their trades.[3] The conjunction of these sell orders arriving nearly simultaneously in the market created the unprecedented opening gap. As the insurers sold and market prices fell, the computer programs of other insurers then triggered further sales causing further declines which in turn caused the first group of insurers to sell even more stock, etc. This in turn generated other sell orders from the same sources and the market experienced the computer-driven meltdown, just as had been predicted a few months earlier by the chairman of the New York Stock Exchange.

---

[1] "The rapid rise in the popularity of portfolio insurance strategies also contributed to the market's rise. Pension fund managers adopting these strategies typically increased the funds' risk exposure by investing more heavily in common stock during this rising market. The rationale was that portfolio insurance would cushion the impact of a market break by allowing them to shift quickly out of stocks." *Report of the Presidential Task Force on Market Mechanisms*, January 1988, page 9.

[2] Even if one believes that portfolio insurance was a major factor in the precipitous decline of the stock market, we believe it is stretching the imagination too far to extend this theory to explain the market rise in the year preceding the crash. To put the Brady Commission claim in perspective, consider that only $60 to $80 billion was being applied to systematic portfolio insurance strategies. Of this, at most $10 billion at the market peak in August 1987 represents additional commitment to stock by insurers over and above what they would have invested if they had followed their original asset allocation approaches. Indeed, many insurers were implicitly less invested in equities than they otherwise would have been because many simply added a futures insurance umbrella while leaving their cash market asset allocation unchanged. In addition, many insurers had restarted their policies (thus reducing their commitment to stocks) during the market rise to "lock in" gains. All in all, the additional investment in equities due to portfolio insurance at the peak of the market must have been less than 1/3 of 1 percent of total market capitalization. It is very difficult to believe that such a small amount of buying power, based as it was on a passive informationless strategy, could push the DJIA up over a period of a year even 50 points, let alone the total actual increase of about 1000 points.

An extension of this point of view takes account of the most significant change in U.S. stock markets in the 1980s: Many large investors (not just portfolio insurers) now trade very large blocks of stock or index futures on short notice while the net liquid assets of floor traders has not kept pace. In large part, the willingness to make large trades is due to reductions in the cost of trading following the May 1, 1975, switch from fixed to negotiated commissions. Here are some interesting statistics. The 1986 annual share trading volume on the NYSE had grown 7-fold since 1975. The 1987 share turnover on the NYSE had reached 73 percent, $3\frac{1}{2}$ times greater than in 1975 and block trades represented 50 percent of traded shares, about 3 times greater than their share in 1975. In addition, the introduction of index futures in 1982 and index options in 1983, as well as new technology-based methods of programmed trading, have increased the ease with which portfolios of stocks can be traded as a group. On the other hand, during this period, the net liquid assets of NYSE specialists, as a percent of the dollar value of NYSE trading volume, had fallen by 1986 to one-third of its 1977 level.[4] In this new environment, it becomes easier for exchanges to become overwhelmed with orders on one side of the market.

Plausible as these two theories of the crash sound—one based on a market panic and the other based on large trader transactions—as we search through the wreckage, we come across other evidence that is difficult to reconcile. Here is a brief list: the crash was international in scope; the U.S. stock market continues to experience a significant increase in volatility, even at a six-month remove; and bid-ask spreads in the stock market remain much greater than before the crash. Most jarring, the U.S. market remained within 1.3 percent of its Black Monday close one week after the crash and has still, as this is written six months later on April 18, 1988, not significantly rebounded. If the crash were caused by either of our two explanations, many economists would have expected more of a correction back to pre-crash conditions by now.

---

[3] "The activities of a small number of aggressive trading-oriented institutions...posed the prospect of further selling pressure on Monday. These traders could well understand the strategies of the portfolio insurers and mutual funds. They could anticipate the selling those institutions would have to do in reaction to the market's decline. They could see those institutions falling behind in their selling programs. The situation presented an opportunity for these traders to sell in anticipation of the forced selling by portfolio insurers and mutual funds, with the prospect of repurchasing at lower prices." *Report of the Presidential Task Force on Market Mechanisms*, January 1988, page 29.

[4] For example, the specialist in IBM has capital of $20 million, representing less than .2 percent of IBM's total stock value of $13 billion. This is well in excess of the revised post-crash requirements imposed by the NYSE which would only require, in the case of IBM, a holding of .014 percent.

In a recent UCLA working paper entitled "Portfolio Insurance and Financial Market Equilibrium," Michael Brennan and Eduardo Schwartz start with a "standard" financial market equilibrium model, inject portfolio insurance investors, and ask, given realistic parameter estimates, how much market volatility would be affected. The standard model is a single-period pure exchange economy with continuous trading during the period but consumption only at the beginning and end. Prices are determined by a representative rational risk averse investor with an additive utility function in current and future consumption who assumes his decisions have no influence on prices. After subtracting current consumption, expectations about end-of-period aggregate wealth evolve according to geometric Brownian motion according to a known drift and variance. A representative portfolio insurer is added to this structure and treated as a "pure automaton" who blindly follows a portfolio strategy, known to the other market participants, which yields an insured position on the market portfolio.

Brennan and Schwartz show that for levels of portfolio insurance as high as 5 percent of total market capitalization, while portfolio insurance trades do increase market volatility, the additional volatility is negligible. The actual amount of assets under systematic portfolio insurance at the end of the second quarter of 1987 was between $60 and $80 billion dollars, which compares to a total U.S. equity market capitalization of $3.6 trillion, or about 2 percent. Compared to total market capitalization, including corporate bonds, non-corporate real estate, and other assets, this percentage would be less than .5 percent. That portfolio insurance or even large trades from other investors could create a single-day 20 percent equity market decline in the absence of significant negative news, even without the other surrounding related events we have mentioned, is wildly at variance with the predictions of standard financial models of equilibrium. To be specific, in the absence of significant news, sales by portfolio insurers should be matched by purchases by other investors who are willing to bear more risk to take advantage of the higher expected returns from somewhat reduced prices. On October 19, it is estimated that portfolio insurance trades in S&P 500 index futures and NYSE stocks amounted to only $6 billion out of a total of $42 billion. In contrast, in the standard model, few other investors would have been selling along with the portfolio insurers. Had this happened, October 19th would have probably been just an ordinary day in stock market history.

If we are to believe the market panic theory or the Brady Commission's theory that the crash was primarily caused by a few large traders, we must strongly reject the standard model. That model is based on three important assumptions: (1) rational investor expectations, (2) continuously functioning markets, and (3) continual optimization by all investors.

With rational expectations, in which all investors are fully aware of the strategies followed by portfolio insurers, smart institutional traders could not have expected to profit from front-running. Contrary to rational expectations, investors may have overreacted because they were uncertain about the amount of trading that would be generated by portfolio insurance. Or they may have interpreted the portfolio insurance trades as information-based, when instead these trades were merely reactive to changes in market prices.

Instead of continuously functioning markets, large sales created an imbalance between supply and demand which led to temporary stock-by-stock closures. This made investors fear that the market mechanism would fail entailing bankruptcies, clearing firm failures, and a total market closure. Understandably, panic could have ensued.

Finally, and most important, contrary to our models, most "investors" have other things on their minds besides the securities markets. Most investors have better things to do; they work during the day and spend time with their families in the evening. Most investors don't know what stocks are worth, tacitly leaving this valuation up to a small set of professional investors. But only a subset of these stand ready to make active investment decisions. All this means that for the purpose of explaining stock price movements, the stock market should be treated as much smaller than the $3.6 trillion quoted earlier.

The visible breakdown of these three assumptions on October 19th may also help explain the related events surrounding the crash. The U.S. market did not rebound because there may be only a few truly "valued-based" investors who are willing to put their money down. Most investors are on the sidelines, implicitly confessing incompetence when it comes to security valuation, themselves confused about whether the crash could reflect a fundamental weakness in the economy. To the extent stocks take value from their perceived liquidity and to the extent the events of the week of October 19th revealed that the market was not as liquid as people believed, stocks are not worth as much today as they were prior to the crash. With thinner markets and with heightened uncertainty stemming from concern over rational expectations and the viability of market mechanisms, price volatility and bid-ask spreads have remained much higher than prior to the crash.

While it is difficult to distinguish between the market panic or large trader theories (and they are hardly mutually exclusive), we believe the evidence favors emphasis of the former. The crash of 1987 was not the first. Many nonequity markets, as well as the U.S. stock market itself on previous occasions and all major foreign stock markets during October, have crashed without investors following systematic portfolio insurance. In many of those cases, a cascade of stop-loss orders, an informal type of portfolio insurance, was a contributing factor. To place systematic portfolio insurance in perspective, an October 19, portfolio insurance sales represented only .2 percent of total U.S. stock market capitalization. Could sales of 1 in every 500 shares lead to a decline of 20 percent in the market? This would imply a demand elasticity of .01—virtually zero—for a market often claimed to be one of the most liquid in the world.

Acceptance of the market panic or large trader theories has important implications for both market regulation and the standard model. The sources of breakdown in the standard model each point clearly in the direction of corresponding regulatory mechanisms. Problems deriving from irrational expectations suggest improvements such as formalized "sunshine" trading which permit investors to distinguish between information-motivated and informationless trades.[5] The potential of discontinuously functioning markets suggests supplementing our current market-making systems with more frequent single-price auctions with open limit order books. Inertia by most investors holding stocks suggests mechanisms which enlarge the scope of the market to involve greater numbers of investors by improving coordination between underlying and derivative markets and increasing global market integration. Unfortunately, regulatory discussion has centered around "circuit breakers" which are largely at cross-purposes with these objectives.

The standard model must itself be seriously flawed. Even if we can only detect its failure during rare events like October 19th, these events make us suspect the model even during normal times. In addition, enough money changes hands during these rare events that they become much more important in the long run than their frequency alone would indicate. We need to build models of financial equilibrium which are more sensitive to real life trading mechanisms, which account more realistically for the formation of expectations, and which recognize that, at any one time, there is a limited pool of investors available with the ability to evaluate stocks and take appropriate action in the market.

---

[5] In a sunshine trade, an investor attempts to preannounce his trading intentions (his identity, order size and timing) several hours prior to the actual trade hoping to deepen the market during the time his trade takes place. Presumably, only non-information based orders would be filled in this way. To formalize this, an exchange would allow posting of sunshine trades on bulletin boards or computer screens on the exchange floor and notice of sunshine trading intentions would be carried via computer to broker-dealers around the world.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 51–72

Lessons on Monetary Policy from the 1980s

Benjamin M. Friedman

The half-decade running from mid-1982 to mid-1987 was a pretty good era for U.S. monetary policy, as these things go. A sharp easing of policy, beginning some time around midyear 1982, helped set in motion a recovery from the most severe business downturn the United States had experienced since the 1930s—a downturn that tight monetary policy earlier on had deliberately brought about in order to slow the alarming acceleration of prices. The recovery that ensued developed into a sustained expansion that continued without interruption through the end of 1987, thereby setting a new record for the longest recorded business expansion in U.S. peacetime experience. A fiscal policy based on unprecedentedly large (and continually growing) structural budget deficits was a major factor underlying this record-length expansion, but at least monetary policy did not stand in the way. Just as importantly, the severe 1981–82 recession served its intended purpose of substantially restoring price stability, and even after five years of expansion inflation remained modest by recent historical standards. At least as judged by these outcomes for the standard objectives of macroeconomic policy, U.S. monetary policy was a distinct success during these years.

Economists hoping to say something useful about monetary policy have had a tougher time. The quantitative relationships connecting income and price movements to the growth of familiar monetary aggregates, including especially the $M1$ measure of the money stock that had been the chief focus of monetary policy during 1979–82, utterly fell apart during this period.[1] Moreover, the collapse of these long-standing

---

[1] The $M1$ money stock consists of currency, checkable deposits (including both non-interest-bearing demand deposits and interest-bearing NOW accounts) and travelers checks; its value as of June 1987 was $747

■ *Benjamin M. Friedman is Professor of Economics, Harvard University, Cambridge, Massachusetts.*

empirical regularities was not merely a matter of larger than usual quarter-to-quarter or year-to-year variances around longer-run benchmarks that otherwise continued to be reliable. Double-digit *M*1 growth, sustained on average over fully five years, repeatedly led prominent economists who had relied on these relationships in the past to offer widely publicized warnings of an imminent re-acceleration of prices. Yet the inflation rate fell dramatically, and then remained low. The presumption that "inflation is always and everywhere a monetary phenomenon" became progressively less compelling as a substantive rather than merely tautological description of the determination of prices.

Economists who preferred to think about monetary policy in different terms had no more success in fitting the major developments of this period into some alternative conceptual framework. Relationships connecting income and prices to the monetary base, or to measures of credit, fell apart just as visibly as did those centered on *M*1. Although real interest rates declined somewhat from the record levels posted in 1981 and early 1982, they nevertheless remained historically high throughout the next five years. Yet the economy's rate of growth in after-inflation dollars was about average for post war business cycle expansions, and in this case the expansion continued on for five years without even a single negative quarter.

In the meanwhile, the rapidly changing structure of the U.S. international economic position brought importance of a whole new magnitude to complications that most economists addressing questions about U.S. monetary policy had acknowledged often enough in the past, but nonetheless had usually ignored in substance. The U.S. current account balance, which had frequently changed sign since World War II but had always remained trivially small in comparison to the economy's size, suddenly became large enough to matter in a domestic macroeconomic context. The dollar exchange rate therefore emerged as a primary focus of U.S. macroeconomic policy. In addition, just within these few years the steady string of mounting current account deficits transformed the United States from the world's largest creditor country to the largest debtor. The exchange rate therefore acquired a new importance in the U.S. financial markets as well.

In the eyes of many economists, the Federal Reserve System has been steering without a rudder ever since it effectively abandoned its commitment to monetary growth targets in 1982. The visible success of monetary policy during the past half-decade is therefore all the more puzzling. In fact, over the course of this period the Federal Reserve's conduct of monetary policy appears to have centered ever more

---

billion. *M*2 consists of *M*1 plus a much larger quantity of savings-type accounts, including ordinary passbook accounts and certificates of deposit (in amounts up to $100,000), money market deposit accounts and money market mutual funds (both of which can have limited checking facilities), and overnight repurchase agreements and Eurodollars; its June 1987 value was $2.8 trillion. *M*3 consists of *M*2 plus institutionally oriented instruments like certificates of deposit in amounts over $10,000 and money market mutual funds used by institutional investors, as well as repurchase agreements and Eurodollars extending beyond overnight; its June 1987 value was $3.6 trillion.

closely on controlling short-term interest rates. Whether this development really means that U.S. monetary policy has now returned to the conceptual basis of a quarter-century ago, when short-term nominal interest rates (or their equivalent) were practically the only focus of the policy process, remains unclear. At the very least, it raises the issue of whether the blatant flaws that crippled this policy strategy in the past were inherent and unavoidable. The other side of the same question is what will happen if the Federal Reserve continues to pursue what is basically an interest rate strategy, if and when inflation again becomes a major problem.

The 1979–82 Experiment and Its Aftermath

The basic facts describing the conduct of U.S. monetary policy in recent years are well known, although their interpretation has been the subject of much disagreement. In October 1979, the Federal Reserve System announced that it was adopting a new policy strategy placing primary emphasis on reducing over time the growth of the money stock. Further, to gain better control over money growth the Federal Reserve would be implementing new operating procedures, based in the first instance on the stock of nonborrowed reserves rather than the federal funds rate or some other short-term interest rate. The bands within which the federal funds rate would be free to fluctuate in the interval between meetings of the Federal Open Market Committee were accordingly widened by more than an order of magnitude. Several more narrowly technical measures, designed in principle to enhance control over money growth, were also part of the overall package.

The immediate motivation underlying this dramatic move was the rapidly deteriorating inflation situation, together with growing concerns about the dollar exchange rate. The overall U.S. inflation rate, as measured by the GNP deflator, had risen from a post-recession low of 6.4 percent in 1976 to 8.9 percent in 1979. The comparable rise measured by consumer prices was from 4.8 percent to 13.3 percent, the highest rate since 1946. At the same time, dollar exchange rates declined between 1976 and 1979 by 16.4 percent on a trade-weighted basis (14.5 percent after correcting for differential inflation). Especially once the OPEC cartel announced yet another major hike in crude petroleum prices, fears of an uncontrollable inflation spiral or a precipitous decline in the dollar, or both, began to spread.

Although the Federal Open Market Committee had at least mentioned money growth targets in its formal policy directives ever since 1970, and had formally reported money growth targets to Congress since 1975, in practice there was little correspondence between the stated targets and actual money growth. Evidence from the 1970s shows that the Federal Reserve did systematically adjust the federal funds rate in the direction required to offset deviations of actual money growth from the targets, but that the magnitude of these adjustments was far too small to be effective for plausible estimates of the interest elasticity of money demand.[2] Perhaps more

[2] See, for example, De Rosa and Stern (1977) and Lombra and Moran (1980).

importantly, the Committee typically did not set the coming year's targets so as to make up for past deviations, but instead let bygones be bygones and so treated each year as independent of the past. As the years rolled on, the difference between actual and targeted money growth was usually positive, and the average difference was positive. As a result, the actual rate of *M*1 growth gradually drifted upward, from 4.9 percent per annum during 1965–69 to 6.1 percent per annum during 1970–74, and then 7.1 percent per annum during 1975–79 (although the fastest growth for any year during this period, 8.4 percent, occurred during 1972).

The effects of the new combination of policy strategy and policy tactics implemented in October 1979 were immediately visible, and they continued to be so for the next several years, although in some aspects they ran counter to the new policy's declared intent. At the most basic level, the Federal Reserve did carry through on its commitment to contain, and then reduce, the growth rate of the *M*1 money stock. The rate of *M*1 growth, measured from the fourth quarter of one year to the fourth quarter of the next (the same basis used for officially reporting money growth targets to Congress) had been 8.2 percent in 1978. The "gradualist" objective of reducing the money growth rate by one percent each year, until it reached a level consistent with price stability, would imply targets of roughly 7 percent for 1979, 6 percent for 1980, 5 percent for 1981, and so on.[3] Although *M*1 had grown at a 9.0 percent per annum rate in the first three quarters of 1979, with the sharp policy shift in the final quarter the growth for 1979 as a whole was 7.9 percent. A continuation of the new policy delivered 7.3 percent *M*1 growth in 1980, and 5.1 percent in 1981 (see Figure 1). For this three-year period viewed as a whole, the new policy did manage to achieve results roughly consistent with the objective of reducing the money growth rate by one percent per year.

Not surprisingly, delivering on this objective involved interest rates that were both higher on average and also more volatile.[4] Nominal interest rates immediately rose to record highs, and then declined sharply as the economy entered a recession and the Federal Reserve Board also imposed credit controls as authorized by President Carter under the Credit Control Act of 1969.[5] Nominal interest rates then rose to yet new record highs in 1981 after business began to recover and credit controls were no longer in effect. Interest rates were high in real terms as well. The difference between the three-month Treasury bill rate and the next quarter's inflation rate, which had

---

[3] The official *M*1 target range for 1979 was 5-8 percent.

[4] Increased short-run volatility of short-term interest rates, as a result of no longer accommodating temporary disturbances affecting money demand, is a straightforward implication of Poole's (1970) analysis of the money growth target strategy for monetary policy. Whether long-term interest rates should be expected to be more or less volatile is a more complicated question, however, involving changing risk factors and expectations of future inflation and interest rates.

[5] The 1969 legislation under which the Board acted was quite far reaching, empowering the Federal Reserve Board, whenever explicitly authorized by the President, to "prohibit or limit any extensions of credit under any circumstances the Board deems appropriate." In 1980 the Board proceeded under this authority to impose special reserve-type requirements on increases in certain kinds of consumer credit by all lenders (including non-banks), on increases in deposits at money market mutual funds, and on increases in non-deposit liabilities at banks that were not members of the Federal Reserve System.

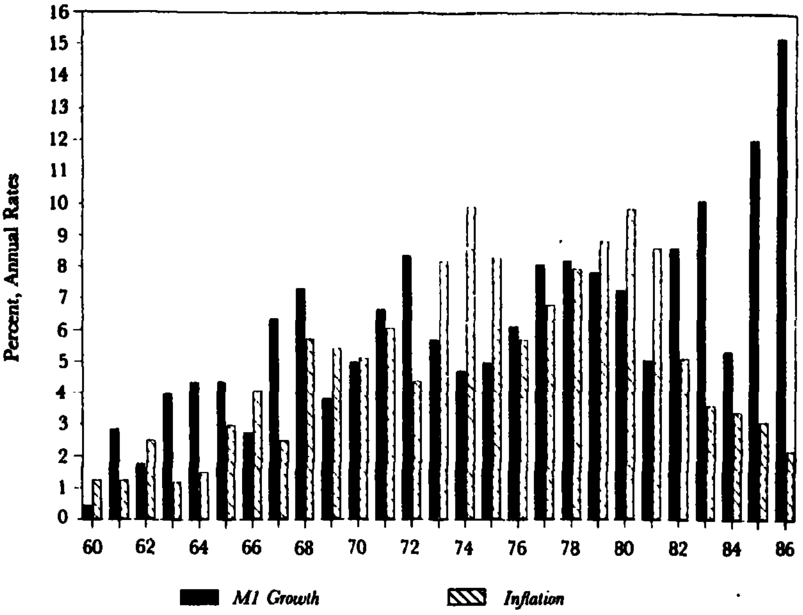


Fig. 1. Money growth and price inflation

averaged close to zero during the post World War II period up until then, fluctuated in the 4-8 percent per annum rage throughout 1981 and 1982. Short-run volatility of interest rates—month-to-month, day-to-day, and even within the trading day—increased by what for most measures were large multiples.

The most striking respect in which the results of the new monetary policy did not correspond to its declared intent was that the short-run volatility of money growth increased sharply too. The Federal Reserve's success in gradually reducing the yearly *M*1 growth rate—8.2 percent in 1978, 7.9 percent in 1979, 7.3 percent in 1980, 5.1 percent in 1981—did not carry over to the quarterly growth rate, which varied in the immediate post-1979 period from a high of 21.1 percent per annum in 1980:*Q*3 to a low of *minus* 5.8 percent per annum in 1980:*Q*2, nor to any finer time calibration. The standard deviation of *M*1 growth from one quarter to the next jumped from 2.37 percent per annum during 1970:*Q*1–1979:*Q*3 to 6.64 percent per annum during 1979:*Q*4–1982:*Q*2 (although omitting the two back-to-back extreme quarters during the 1980 credit control episode, the increase was only to 3.00 percent per annum). Although the Federal Reserve had never formally acknowledged any commitment to stabilize money growth on a within-year basis, critics of monetary policy, both at the time and subsequently, focused on this increase in short-run volatility as casting doubt on the strength of the central bank's commitment to money growth targets in a broader perspective.

The monetary policy experiment that commenced in October 1979 ended almost as abruptly as it had begun. By the summer of 1982 the slowdown in U.S. business activity had developed into what was, by many measures, the most severe recession since the great depression of the 1930s. Unemployed labor, idle industrial capacity and business bankruptcies were all at post war record levels. At the same time, bank failures had also reached record levels, and there were increasing signs of fragility throughout the financial structure. In August 1982 an impasse over Mexico's ability to meet its dollar-denominated external obligations called widespread attention for the first item to the debt problems not just of Mexico but of developing countries throughout Latin America and sub-Saharan Africa.

Yet the demand for money balances continued to be strong. Following the 5.1 percent growth in 1981, the Federal Reserve had set a 2.5-5.5 percent target range for *M*1 growth in 1982. The revised data now available show that actual *M*1 growth was within this range (4.6 percent per annum) during the first half of 1982; but the data available at that time showed *M*1 growth running consistently above the stated target range, despite continuing high interest rates and contracting real economic activity. In August the Federal Reserve allowed short-term interest rates to drop by almost three percentage points notwithstanding the emergence of still more rapid money growth. In October, within less than a week of the third anniversary of the announcement that had proclaimed the new dedication to money growth targets, Federal Reserve Chairman Paul Volcker publicly acknowledged that the *M*1 growth target was no longer in effect. With a further speed-up in the fourth quarter, *M*1 growth for the year 1982 came to 8.6 percent—above the target range, above the 1978 growth rate from which the program of cutting back on money growth had originally begun, and even above the postwar record growth rate set back in 1972.

The experience of the next five years largely continued the course set in the latter half of 1982. The yearly *M*1 growth rate was in double digits in each of 1983, 1985 and 1986 (see again Figure 1). For the five-year span ending at mid-1987, the average *M*1 growth was 10.8 percent per annum. Only in 1984 did actual money growth fall within the stated target range. In both 1983 and 1985 the Federal Reserve officially changed the target range at midyear, once the difference between actual and targeted *M*1 growth became obvious. In 1986 the Federal Reserve suspended its *M*1 target range without setting a new one. In 1987 the Federal Reserve simply set no target range for *M*1 growth at all.

Throughout the post-1982 period the Federal Reserve's official pronouncements continued to emphasize targets for broader monetary aggregates in place of *M*1, but it is not clear to what extent these measures genuinely guided monetary policy. Actual *M*2 growth did fall within the stated target range in every year between 1983 and 1986.[6] Actual *M*3 growth exceeded the stated target range in 1983 and 1984, but fell within it in 1985 and 1986. In 1987, until the October drop in stock prices, the Federal Reserve either instigated or accepted (depending upon one's perspective) rising market interest rates, and also raised the discount rate, despite the fact that both *M*2 growth and *M*3 growth were falling short of the stated target range; and the decline of market interest rates immediately after the stock market crash appeared to have little to do with money growth patterns in any direct way.

---

[6] In 1983 the stated target range for *M*2 growth covered only part of the year.

By contrast, short-term nominal interest rates since 1982 have resumed the smooth pattern characteristic of the pre-1979 era, thereby suggesting a renewed role for interest rates—as before—at the center of the monetary policymaking process. The standard deviation of the month-to-month change in the three-month U.S. Treasury bill rate, for example, had risen from .42 percent per annum between January 1970 and September 1979, to 1.54 percent per annum between October 1979 and September 1982. It dropped back to .32 percent per annum between October 1982 and June 1987. The standard deviation of the month-to-month change in the federal funds rate rose from .50 percent to 1.92 percent, and then fell to .38 percent, over the same three periods.

Although some of this return to interest rate smoothness on a month-to-month basis may simply have reflected the continuity of the business expansion and the absence of a re-acceleration of prices during this period, the pronounced stability of short-term interest rates over a substantial span of time more likely indicates the return to a monetary policy approach based on closely controlling interest rate movements. At the least, it presents a strong contrast to the behavior that the Federal Reserve has accepted for the monetary aggregates.

Collapse of the Money-Income and Money-Price Relationships

What makes this unusual record of monetary policy actions look so successful in retrospect is that the bizarre behavior of money growth in no way corresponded to the behavior of income or prices. The familiar relationships that had characterized prior experience simply disappeared.

Instabilities in the money-income relationship—or, in more sophisticated forms, the money demand function—had actually begun to become more pronounced as early as the mid-1970s, and their appearance had already spawned a substantial new body of empirical literature even before the new monetary policy experiment commenced in October 1979.[7] By 1980 the Federal Reserve System had already adopted a whole new set of definitions of the monetary aggregates, designed in part to overcome just such difficulties. A survey paper bearing the suggestive title "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," and including more than eighty references, was already in print in the *Journal of Economic Literature* before Paul Volcker acknowledged in October 1982 that the Federal Reserve was suspending its $M1$ growth target (Judd and Scadding, 1982.)

As Figure 2 makes clear, however, the instability that generated so much concern and research in the pre-1982 period was small stuff in comparison to what followed.

---

[7] The standard reference to state first is the contrast between the findings in Goldfeld (1973) and Goldfeld (1976). The most widely read studies done at the time by the Federal Reserve's own staff include Enzler *et al.* (1976), Porter *et al.* (1979), and Simpson and Porter (1980).

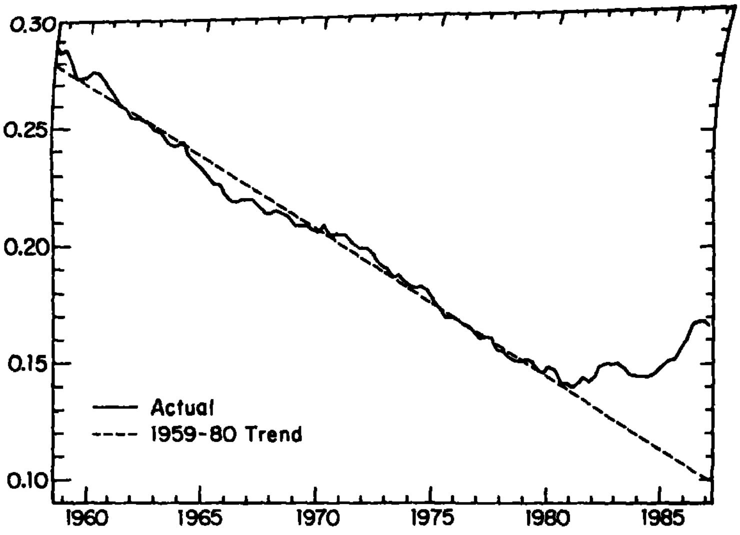


Fig. 2. Ratio of money to GNP

Note: The series plotted is the ratio of $M1$, as a quarterly average of monthly data (source: Board of Governors of the Federal Reserve System) to quarterly GNP at annual rates (source: U.S. Department of Commerce). Both series are seasonally adjusted.

The figure plots the ratio of the $M1$ money stock to GNP for each quarter since the start of the redefined $M1$ series in 1959:$Q1$. Through the end of 1980, the $M1$-to-GNP ratio displayed the familiar downward trend of roughly 3 percent per annum that most students of the money-income relationship had come to see as inevitable in the post-war period, with a standard deviation around this trend of only .0044 (in comparison to a 1980:$Q4$ value of .1466). After 1980 the $M1$-to-GNP ratio not only experienced wider fluctuations but even reversed course. A simple extrapolation of the 1959-80 trend implies a ratio of .1007 by 1987:$Q2$ (the last quarter plotted). The actual value in 1987:$Q2$ was .1686, different from the trend extrapolation by more than 15 times the 1959-80 standard deviation.

Discussion of this phenomenon at the popular level has typically offered as an explanation the fact that "velocity" has declined. Because the so-called income velocity of money is nothing other than the ratio of GNP to money (the reciprocal of the ratio plotted in Figure 2), however, such explanations are completely empty of content. Given the definition of "velocity" in this context, the fact that velocity declined is simply identical to the fact that money grew rapidly while income did not. Saying that money growth outpaced income growth because velocity declined is like saying that the sun rose because it was morning.[8]

---

[8] I owe the analogy to William Bennett.

The mere fact of instability in the simple money-income ratio need not, of course, imply instability in more fully specified behavioral representations of the money-income relationship. The impression that stands out on a first glance at Figure 2 is representative of the results that researchers employing a variety of statistical strategies have found, however. A standard Goldfeld-type money demand function, estimated for quarterly data spanning 1952:$Q$3–1979:$Q$3, indicates a standard error of .42 percent. Extending the sample to 1986:$Q$4 raises the standard error to .61 percent. Deleting the earlier data, so that the sample is 1974:$Q$2–1986:$Q$4, further raises the standard error to .84 percent. Dynamic out-of-sample simulations of such equations deliver cumulative errors with root mean squares in the range of 4-8 percent for different parts of the post-1982 period, in comparison with 0.5–1.5 percent for different parts of the pre-1974 period. Attempts to do better with alternative specifications have met at best only very limited success.[9]

The story is approximately the same for efforts to investigate the money-income relationship from the perspective of determining income rather than money. A "St. Louis" type equation relating nominal GNP to four-quarter lags on both $M1$ and high-employment government expenditures, estimated in logarithmic differences for quarterly data spanning 1960:$Q$1–1979:$Q$3, indicates an adjusted coefficient of determination ($\bar{R}^2$) of .32. Extending the sample to 1986:$Q$4 reduces the $\bar{R}^2$ to .11. Deleting the earlier data, so that the sample is 1970:$Q$3-1986:$Q$4, further reduces the $\bar{R}^2$ to just .02. More sophisticated autoregression methods testing for a significant role of money in "causing" either nominal or real income, in the sense of accounting for income fluctuations not already accounted for by prior fluctuations in income itself, have produced results that are sufficiently varied to generate more skepticism than confidence in any strong conclusion on the subject, either positive or negative.[10]

Finally, in considering the money-*price* relationship it is even necessary to be on guard against results that are strongly statistically significant but with the wrong sign to make any sense in economic terms. The double-digit average growth rate maintained for five years following mid-1982 represents the most rapid sustained money growth the United States has experienced since World War II, yet these same years also saw the strongest sustained deceleration of prices in the postwar period (see again Figure 1). Price inflation as measured by the GNP deflator peaked at 9.7 percent in 1981 and declined in each of the next five years, reaching 2.6 percent in 1986. The rate of increase of consumer prices peaked at 13.3 percent in 1979 and declined in all but one of the next seven years, reaching 1.1 percent in 1986. Given the role that high-variance observations play in dominating results based on the least-squares methodology, as of the late 1980s it is necessary to take care not to find results indicating that *faster* money growth implies *slower* inflation.

---

[9] The specific results cited here are from Goldfeld and Sichel (forthcoming), which also provides an extensive survey. Roley (1985) also showed the results of experimenting with a wide variety of alternative specifications.

[10] See, for example, the differing results reported in Friedman (1986), Eichenbaum and Singleton (1986), and Stock and Watson (1987).

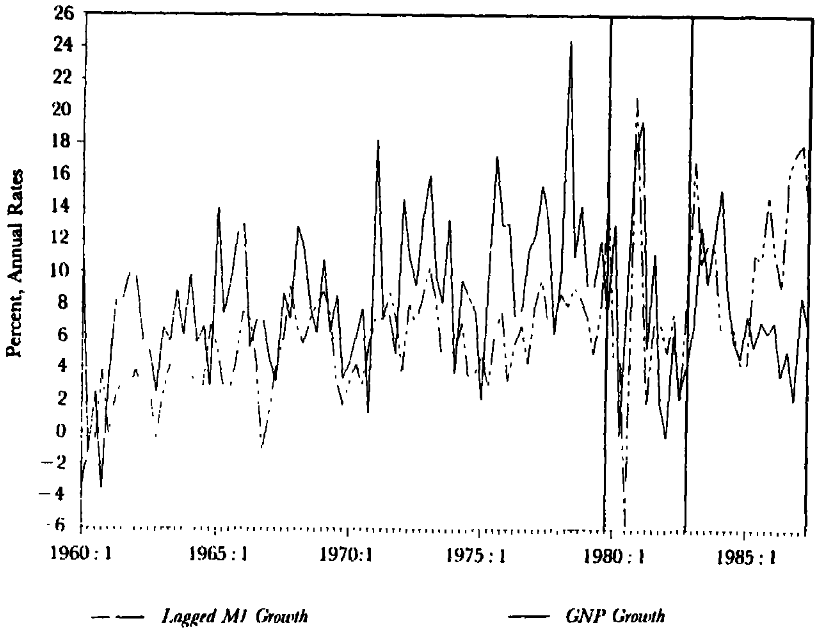


Fig. 3. Growth rates of GNP and lagged money

Two examples, both drawn from the same paper, readily illustrate the pitfalls that confronted anyone who continued to rely closely on straightforward money-income and money-price relationships during this period. First, Figure 3 is an expansion, both backward in time and forward, of a figure included in a paper by Milton Friedman (1984) in the *American Economic Review*. The figure plots the respective annualized quarter-to-quarter growth rates of GNP and, with a one-quarter lag, *M*1. The figure covers 1960:*Q*1–1987:*Q*2 and distinguishes three time intervals. The middle one, 1979:*Q*4–1983:*Q*4, is identical to that plotted by Friedman. It spans the period from the October 1979 inception of the new monetary policy experiment through what was presumably the most recent observation available as of his time of writing.

After pointing out that the correlation between these two series during 1979:*Q*4–1983:*Q*4 was .46, or .71 after eliminating the two quarters affected by the credit control episode, Friedman wrote, "Two things are notable about the relation between money and income in these years: first, the lag is both shorter on the average and less variable than in earlier years, second, the relation is unusually close. I believe that both are a consequence of the exceptionally large fluctuations in *M*1 growth. The effect was to enhance the importance of the monetary changes relative to the numerous other factors affecting nominal income and thereby to speed up and render more consistent the reaction."

Table 1 summarizes the record of the GNP-to-lagged-*M*1 growth correlation and the variability of *M*1 growth for the three intervals shown in Figure 3. Money growth on a quarter-to-quarter basis (as used by Friedman in his paper) was certainly more

Table 1

Money growth volatility and the GNP-to-lagged-*M*1 correlation, 1960–1987

Sample	*Standard deviation of M1 growth*	*Correlation between GNP growth and lagged M1 growth*
1960:*Q*1–1979:*Q*3	2.87%	.47
1979:*Q*4–1983:*Q*4	6.16	.45
1979:*Q*4–1980:*Q*1, 1980:*Q*4–1983:*Q*4	4.18	.47
1984:*Q*1–1987:*Q*2	4.80	–.10

Data are seasonally adjusted at annual rates. Money data are quarterly averages.

variable during 1979:*Q*4–1983:*Q*4 than it had been during the prior two decades. The GNP-to-lagged-*M*1 correlation was not "unusually close" during 1979:*Q*4–1983:*Q*4 compared to the past, however. The correlation of .45 computed over these eleven quarters (Friedman reported .46) is essentially identical to that for the previous 79 quarters. Excluding 1980:*Q*2 and 1980:*Q*3 reduces the variability of money growth, but does not materially affect the GNP-to-lagged money correlation. (Subsequent data revisions have reduced the .71 correlation reported by Friedman to .47 as shown in Table 1—identical to the correlation for the earlier period.)

More importantly, what stands out in both Table 1 and Figure 3 is the changes that occurred after 1983. Although the variability of money growth remained high, the positive GNP-to-lagged-*M*1 correlation disappeared entirely. In its place is a small *negative* correlation.

Table 2, focusing on the money-price relationship, is simply an updated version of a table that Friedman presented in the same 1984 paper. The horizontal line in each column indicates entries not included in the original version. In describing the data shown above the two lines, Friedman wrote, "The long-period evidence suggests that inflation has much inertia and that the lag between money and inflation is of the order of two years. Table [2] shows that this relation has held in recent years as well. There is a one-to-one relation between movements in monetary growth, and in the GNP deflator two years later over successive two-year periods since 1971... The increased rate of monetary growth in the 1981-83 biennium suggests that we have passed the trough in inflation and that inflation will be decidedly higher from 1983 to 1985 than it was from 1981 to 1983."[11]

---

[11] Friedman made the some prediction more forcefully in writings directed at broader audiences. In a column in the September 26, 1983 issue of *Newsweek*, for example, Friedman wrote, "Inflation has not yet accelerated. That will come next year, since it generally takes about two years for monetary acceleration to work its way through to inflation...The monetary explosion from July 1982 to July 1983 leaves no satisfactory way out of our present situation...The result is bound to be renewed stagflation—recession accompanied by *rising* inflation and high interest rates." A lengthy interview in the March 19, 1984, issue of *Fortune* indicated that Friedman "...also sees a strong possibility that by the end of [1984] inflation could reach an annual rate as high as 9%."

Table 2

Rates of change in money and in inflation eight quarters later

	Annual rate of change over eight quarters		
Period for money	*M1*	*Deflator eight quarters later*	*Period for deflator*
1971:$Q$3–1973:$Q$3	6.9%	9.5%	1973:$Q$3–1975:$Q$3
1973:$Q$3–1975:$Q$3	5.2	6.3	1975:$Q$3–1977:$Q$3
1975:$Q$3–1977:$Q$3	6.4	8.3	1977:$Q$3–1979:$Q$3
1977:$Q$3–1979:$Q$3	8.6	9.4	1979:$Q$3–1981:$Q$3
1979:$Q$3–1981:$Q$3	6.1	4.8	1981:$Q$3–1983:$Q$3
1981:$Q$3–1983:$Q$3	9.2	3.3	1983:$Q$3–1985:$Q$3
1983:$Q$3–1985:$Q$3	8.1	2.8	1985:$Q$3–1987:$Q$3

Data are seasonally adjusted.

Source: Friedman (1984), updated. (The entries above the lines differ from Friedman's because of subsequent data revisions, but the differences are slight.)

As the below-the-line entries in Table 2 show, quite the opposite happened. Growth of $M1$ during 1981:$Q$3–1983:$Q$3 was the fastest for any of the six biennia in Friedman's sample, but inflation in 1983:$Q$3–1985:$Q$3 turned out to be the lowest. Rapid money growth continued in 1983:$Q$3–1985:$Q$3, but inflation slowed still further in 1985:$Q$3–1987:$Q$3. The simple correlation between the two time series shown, calculated for the first five observations only, is .70. Calculated for all seven observations, the correlation is *minus* .23.

Other Money and Credit Aggregates

The breakdown of long-standing relationships to income and prices has not been confined to the $M1$ money measure. Neither $M2$ nor $M3$, nor the monetary base, nor the total debt of domestic nonfinancial borrowers has displayed a consistent relationship to nominal income growth or to inflation during this period. On a quarter-to-quarter basis, standard relationships like Goldfeld-type equations fitting movements in these aggregates to movements of income and interest rates, or St. Louis-type equations fitting movements of nominal income to movements of an aggregate and a measure of fiscal policy, showed pronounced deterioration for each of these aggregates. (The *largest* $\bar{R}^2$ for any of these St. Louis equations, estimated for quarterly data spanning 1970:$Q$3–1986:$Q$4, is .09.) On a longer-term basis, the average growth rate for each of these aggregates during the half-decade from mid-1982 to mid-1987 was in excess of any prior postwar experience; yet inflation lessened substantially, and the average growth of nominal income was hardly extraordinary. By mid-1987 the ratio of each aggregate to nominal GNP was above the level implied by an extrapolation of the corresponding pre-1980 trend by an amount ranging from three standard deviations (for *M*2) to twenty-three (for the credit aggregate), based on the pre-1980 variability. It is difficult to imagine how anyone could have successfully predicted the behavior of either income or prices during this period on the basis of foreknowledge of the path of any—or, for that matter, all—of these aggregates.

The manifest failure of the credit aggregate to perform satisfactorily in this context perhaps merits a special comment. I had earlier advocated the use of a broad credit aggregate in conjunction with one or more monetary aggregates in formulating U.S. monetary policy, precisely on the ground that credit might provide some safeguard against false signals given by the monetary aggregates under conditions of instability affecting the public's demand for money (see, for example, Friedman, 1982, 1983). Because credit is a measure of activity on the liability side of the public's balance sheet, while the monetary aggregates are various measures of the non-bank public's assets, it seemed (and to me still seems) reasonable to think that expanding the information base explicitly underlying the monetary policy process to encompass both money and credit measures would provide potentially useful diversification in the context of portfolio behavior that is at best imperfectly understood, and inevitably subject to a multiplicity of shocks. In addition, empirical investigations relying on a variety of statistical methods indicated little basis for concluding that the total debt of all domestic nonfinancial borrowers was any less (or any more) closely related to movements of income or prices than was any of the standard monetary aggregates.[12]

In any event, the movement of credit during the post-1982 period bore no more relation to income or prices than did that of any of the monetary aggregates. Worse still, the false signals provided by the growth of credit were in the same direction as those provided by the growth of money. Figure 4 plots the credit-to-GNP ratio for the same sample for which Figure 2 shows the *M*1-to-GNP ratio. After decades of trendless stability, the credit ratio began an unprecedented climb in 1982 which has not stopped as of the time of writing. Moreover, disaggregated data show that essentially all categories of domestic nonfinancial borrowers—including the federal government, state and local governments, individuals, and businesses—have played major roles in this extraordinary surge of indebtedness.[13] Anyone who had relied on prior credit-based relationships to predict the behavior of income or prices during this period would have made forecasts just as incorrect as those derived from money-based relationships. Anyone who had derived additional confidence in such predictions

---

[12] There are several obvious problems with attempting to measure the relevant concept of credit in this way. One is simply that the available data measure long-term debts at par value rather than at market prices (or some equivalent for nonmarketable debts). Another is that, although the category of "nonfinancial" borrowers excludes any entity explicitly set up as a financial intermediary, there is inevitably some degree of double-counting due to what amounts to financial intermediation carried out by ordinary businesses and even individuals. Whether this problem is more or less severe than comparable problems affecting the monetary aggregates—for example, the apparently widespread use of U.S. currency in black markets around the world, or even in the United States for a variety of purposes not related to familiar theories of demand for money—is an empirical question.

[13] The one exception is the farm sector.

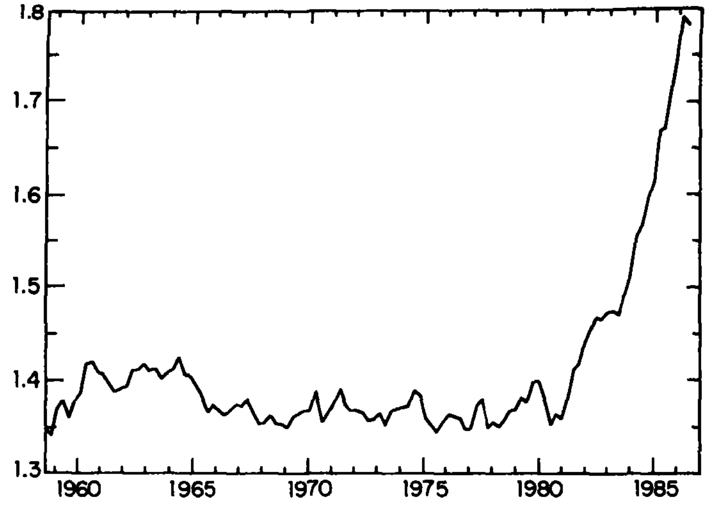


Fig. 4. Ratio of credit to GNP

Note: The data are analogous to those in Figure 2 except that the numerator is end of quarter outstanding debt of domestic nonfinancial borrowers (*Source*. Board of Governors of the Federal Reserve System.)

because the respective signals given by both money and credit confirmed one another would have found that confidence misplaced.

I have speculated elsewhere on the causes of the breakdown of the relationship between credit and income in the 1980s (Friedman, 1987). For purposes of the monetary policy issues under discussion here, it is sufficient to say that attempts to "fix up" this relationship in any simple way are likely to be no more successful than such efforts directed at parallel relationships for the monetary aggregates have been.

A Relationship That Did Hold Up

In sharp contrast to the collapse of relationships connecting the ultimate objectives of monetary policy to standard quantity measures of policy actions, the long-standing relationship between the two most prominent macroeconomic policy objectives—inflation and unemployment—remained intact during this period. The point is of some interest because one of the principal supposed merits widely claimed in favor of the use of publicly announced money growth targets for monetary policy was a potential lessening of the real costs of disinflation. The idea was that public knowledge of such targets would affect expectations in such a way as to minimize (according to some models, to eliminate altogether) the usual negative impact of disinflationary monetary policy on employment, output, incomes, and profits.

What this idea should have led one to predict about the real costs of the major disinflation achieved in the United States in the 1980s is far from clear. For almost three years beginning in October 1979, the Federal Reserve did approximately carry through on its widely publicized intention to reduce the yearly rate of money growth by one percent per year—that is, as long as everyone understood that "money" meant *M*1. Because the familiar claim that most measures of money move roughly together over time turned out to be false, however, anyone who watched *M*2 or *M*3 instead of *M*1 would have seen no evidence of monetary deceleration.[14] Moreover, on a within-year basis even *M*1 growth became not more regular but less after October 1979 (see again Table 1). After mid-1982 there was no reason for anyone to find the Federal Reserve's commitment to its stated money growth targets credible.

Regardless of whether any part of the policy experience of the 1980s constituted a good test for an effect of preannounced money growth targets in reducing the real costs of disinflation, it is clear that no such reduction occurred. Ten years ago, Arthur Okun (1978) surveyed a variety of econometric estimates of these costs, none of which incorporated any expectations effects due to reliance on preannounced money growth targets. Okun's survey indicated that the cost of each percentage point reduction in inflation achieved by monetary policy would be between two and six "point-years" of unemployment, with a median estimate of three point-years.[15] Table 3 shows the annual rate of change of the GNP deflator and the annual average unemployment rate beginning in 1978. The table also shows, for years beginning in 1980, the cumulative excess of the unemployment rate above 6 percent (the approximate average for the two prior years and a standard "full employment" benchmark.

The slowing of inflation from nearly 10 percent per annum in 1980-81 to roughly 3 percent per annum a half-decade later required approximately 14 point-years of unemployment—right at the lower end of the range implied by the econometric models Okun surveyed. Especially in light of the evidence suggesting that something like a third (and perhaps as much as a half) of this disinflation was a product of the 74 percent appreciation of the dollar's trade-weighted exchange rate between 1980 and early 1985—an appreciation which has been almost entirely reversed as of the time of writing—this result seems fully consistent with Okun's prediction.[16]

Everyone had always known that sufficiently tight monetary policy, maintained for a sufficiently long time, could halt even the most deeply rooted inflation. The

---

[14] For the five years 1978–82, the simple correlations among the fourth-quarter-over-fourth-quarter growth rates of the major M's were each *negative*: −.53 between *M*1 and *M*2, −.57 between *M*1 and *M*3, and −.12 between *M*2 and *M*3.

[15] A "point year" of unemployment is one percentage point of unemployment in excess of the rate that corresponds to "full employment," maintained for one year. Some writers—for example, Fischer (1985)—have focused on real output rather than unemployment, and have argued on that basis that the post-1980 disinflation involved smaller costs than Okun's survey implied. The focus of the evidence that Okun surveyed was the inflation-unemployment relationship, however. His translation of the cost estimate into foregone real output simply relied on the usual three-for-one "Okun's Law" relation, which has not held up during the 1980s.

[16] See, for example, Sachs (1985) for an analysis of the importance of the dollar's appreciation in the U.S. disinflation.

Table 3
Percentage rates of inflation and unemployment, 1978–1987

	Inflation rate	*Unemployment rate*	*Cumulative excess unemployment*
1978	7.3%	6.1%	—
1979	8.9	5.8	—
1980	9.0	7.1	1.1%
1981	9.7	7.6	2.7
1982	6.4	9.7	6.4
1983	3.9	9.6	10.0
1984	3.7	7.5	11.5
1985	3.2	7.2	12.7
1986	2.6	7.0	13.7
1987	3.2	6.5	14.2

Data for 1987 based on first half year only.

reluctance to proceed in that fashion lay not in disbelief that such a policy would do its job, but in concern for the resulting real costs. What was new beginning in October 1979 was the willingness to bear those costs.

International Complications

The United States is not a small open economy. U.S. production and U.S. incomes bulk sufficiently large in comparison to the combined economic scale of all countries participating in the current free international trading system that independent forces originating in this country can and do affect economic conditions on a worldwide basis. The U.S. financial markets bulk sufficiently large in comparison to world capital markets to exert analogous effects.

The United States is an open economy, however, and the increasing practical importance of this fact is perhaps the most interesting single development of the 1980s from a monetary policy perspective. There are at least two major aspects of the U.S. economy's openness that now matter, far more than earlier on, for purposes of monetary policy. The more straightforward of the two is simply the demonstrated ability of U.S. exports and U.S. imports to differ by a margin large enough to affect the economy's aggregate output and growth.

The possibility of such an outcome had always been present, of course; but throughout the postwar era, until the 1980s, the U.S. current account was always approximately in balance over periods long enough to even out the ups and downs of the business cycle. The current account was in surplus, albeit by less than $1 billion per annum (0.2 percent of GNP), on average during the 1950s. In the 1960s the average annual surplus was $3 billion (0.4 percent of GNP). In the 1970s the current account was in deficit on average, but by less than $1 billion per year (not even 0.1

percent of GNP). In 1980 there was a $2 billion surplus, compared to overall GNP of $2.7 trillion. It is little wonder that many analysts of U.S. monetary policy during these decades practically ignored potential effects on real economic activity via exchange rate channels.

The 1980s have been starkly different. With a 74 percent appreciation of the dollar on a trade-weighted basis between the end of 1980 and early 1985 (64 percent after allowing for differential inflation rates), together with a host of other factors that may have further compounded the problem, the ability of many U.S. industries to compete in world markets all but collapsed. By 1986 the current account deficit was $141 billion, or 3.3 percent of GNP. On an overall basis, including the trade balance, real GNP in 1986 was up over the 1980 level by 16.5 percent. Excluding the trade balance—that is, focusing on domestic absorption rather than production—the increase over the six years was 23.3 percent. Not surprisingly, exchange rate effects on export and import flows have now emerged as a genuine focus of monetary policy concern, not just a subject for obligatory mention.

The more complex aspect of the newly enhanced importance of the U.S. economy's openness from a monetary policy perspective is an outgrowth of the first. A current account deficit means that a country is borrowing from abroad to finance the excess of its imports over its exports (where both are broadly defined)—or, equivalently, to finance the excess of its domestic investment and government deficit over its domestic saving. Since 1980 U.S. fiscal policy has delivered an unbroken string of unprecedentedly large federal government budget deficits, notwithstanding five years of business expansion beginning in 1983. At the same time, domestic saving has been below average despite the introduction of a variety of supposed saving incentives. As a result, the United States has been borrowing from abroad in record volume even though the share of GNP devoted to net investment in both business and residential capital has been unusually small. Especially since 1984, the United States has borrowed from abroad on a larger scale in relation to the economy's size than the country did at any of the four peak periods of reliance on foreign capital to finance its canals, then its railroads, then its steel and other industries, and then its public utilities, during the nineteenth century.[17]

This massive borrowing from abroad transformed the United States from the world's largest creditor country to its largest debtor in just four years. The U.S. net foreign asset position peaked at $141 billion at the end of 1981. It was still $136 billion at the end of 1982. The United States officially became a net debtor again (for the first time since before World War I) in early 1985. By the end of 1986 the U.S. net foreign debt was $265 billion—more than that of Mexico, Brazil and Argentina combined.[18] On a gross basis (that is, without any balancing of U.S. holdings abroad)

---

[17] See Solomon (1986) for a review of the nineteenth century experience.

[18] The official accounts include numerous obvious mismeasurements, but there is no ground for claiming that their sum is very different from zero. The largest adjustments in favor of the United States in a set of true accounts would be the revaluation of U.S. gold stocks, and of the net of U.S. direct investment abroad and foreign direct investment in the United States, to current market values. The largest adjustments against the United States would be the revaluation to market value of U.S. banks' loans to developing countries, and allowance for the accumulation over time of each year's "errors and omissions" flow.

foreign holdings in the United States as of year-end 1986 came to $1.3 trillion, including such readily liquefiable assets as $267 billion of U.S Treasury securities, $149 billion of corporate bonds and other debt securities, $167 of corporate equities, and $477 billion of bank deposits.[19]

As a result of this surge in foreign holdings during the 1980s, foreign investors have assumed an importance in the U.S. financial markets that was previously unknown in modern times. Whether or not foreign investors will bid in volume on any given day's auction of U.S. Treasury securities is now often the paramount focus of attention among participants not just in the U.S. bond market but in the stock market too. Popular discussions frequently raise the fear that foreign investors might decide to "get out" of dollar holdings, with disastrous consequences for the dollar exchange rate as well as for the prices of dollar denominated securities. The notion that foreign investors in the aggregate can sell their dollar holdings to anyone but each other is mistaken, of course, but their attempts to do so can move both exchange rates and securities prices. In particular, given the size of foreign holdings, a widespread move to liquidate bond and equity portfolios as a preliminary to selling dollars (even if only to other foreign buyers) could easily move securities prices by a substantial margin. As a result, managing dollar exchange rates to prevent such occurrences has added a whole new dimension of U.S. monetary policy.

The contrast between the rigid hands-off attitude toward the foreign exchange markets that the Reagan administration espoused during its first four years in office, and the series of ad hoc international agreements that it has orchestrated beginning with the widely publicized meeting at New York's Plaza Hotel in September 1985, is just the most visible part of the increased importance of the exchange rate for the United States. Although there is little hard evidence as yet, it appears that the change has affected U.S. monetary policy in more ongoing, and more subtle, ways as well. For example, even at the simplest level it is no longer so obvious that an open market purchase which lowers short-term interest rates will necessarily lower long-term interest rates as well. If dollar exchange rates fall as a result of lower U.S. short-term rates, foreign investors may sell their long-term U.S. assets as a preliminary to reducing (again, not collectively) their overall dollar holdings, and domestic market participants may seek to do the same in anticipation of the action by foreign investors.[20]

If the net result of these developments amounts to an erosion of sovereignty over U.S. monetary policy, that loss is no more than the inevitable price of becoming a debtor country—and the world's largest, at that. Losing control over one's affairs is part of what a mounting debt level is all about, no less for a borrowing nation than for business or individual borrowers. The policy actions and public discussions that ensued after the October 1987 stock market crash dramatically illustrated the tension that can arise under such circumstances between a monetary policy designed to support the currency and a monetary policy designed to prop up the domestic economy. At least thus far the United States' situation in this regard is not yet so constraining as, for example, Britain's was during the 1950s and 1960s, when repeated ad hoc concessions were necessary to avert the eruption of a "sterling crisis." Even so, the problem of formulating and implementing U.S. monetary policy is now more complicated by one additional dimension.

---

[19] These data are from Scholl (1987), Table 1.

[20] The effect is analogous to what is sometimes claimed along the lines that open market purchases would lead investors to sell long-term bonds out of fear that the resulting increase in money growth would bring higher inflation. Market experience in the United States has not borne out this earlier line of reasoning, but there appears to be more evidence to support the effect operating via exchange rates.

Vacuum at the Center?

It is difficult to escape the conclusion that there is now a conceptual vacum at the center of the U.S. monetary policymaking process. The seemingly endless quest from various quarters to impose some kind of simple rule on the conduct of monetary policy is, at least for now, no longer a going concern. The interactions among money, income and prices during the 1980s—including the half-decade that followed the monetary targets experiment, even more than what happened during 1979–82—have undermined it both intellectually and practically. The claim that reliance on some kind of simple rule would open the way to costless disinflation is at best unproven, and for practical purposes doubtful. The notion that some kind of simple rule would adequately encompass the new complications due to the increased practical importance of the U.S. economy's openness is dubious at best. In retrospect, the contemptuous dismissal by some economists of the Federal Reserve's reluctance to adopt the monetary policy rules they had proposed—and the readily voiced assumption that that reluctance could have stemmed only from ignorance or a faulty set of objectives on the part of Federal Reserve officials, or perhaps even their self-aggrandizement at the public expense—stands as a sorry reminder that economists outside government can also make each of these errors.

The most powerful element of what has happened in this regard in the 1980s is the collapse of the longer-run relationship between money on the one side and income and prices on the other. In this decade the main event has been very different from the quarter-to-quarter or even year-to-year irregularity that was always the focus of debate about whether these relationships were stable before. Proponents of simple monetary policy rules in the past could and did claim that such rules failed to offset short-run economic fluctuations that policy could probably not hope to eliminate anyway, but that over longer horizons the anchor they provided would keep the economy on a steadier course than an alternative policy that attempted to achieve "fine tuning." With nominal GNP by mid-1987 more than 40 percent below the value implied by the long-run relationship to $M1$ which prevailed during 1959–80, the problem is no longer in the fine tuning but in the anchor itself.

The relevant issue here for policy purposes is not whether for any time interval there exists some abstract notion of "money" that conceptually bears a stable and reliable relation to income and prices, and that statisticians can seek to uncover after that interval ends, but whether policymakers can identify and measure that quantity substantially in advance of their need to base planning and operating decisions on it. Appeals to the tradition of the "quantity theory" are of no use in this context in the absence of a clear statement of what is the quantity and what is the theory. Especially in a world of institutions that increasingly blur the distinction between transactions balances and saving balances, being precise about either the theory or the quantity is ever more difficult. The continuing (indeed increasing) interest within the economics profession in some kind of constitutional constraint on monetary policy jars harshly againt the likelihood of what such constraints suggested (but not enacted) in years past would have meant in the 1980s.

Nor is the problem merely one of money (or credit) "targets." The more flexible idea that such variables as money and credit, which are endogenous to the monetary policy process in the short run, should be used as "information variables" to guide the initial setting and ultimate readjustment of whatever genuinely exogenous instrument the central bank is using, has always seemed highly attractive—at least to me. The information-variable approach to monetary policy makes no sense, however, in the absence of a reasonably compact set of variables that reliably provide information about the macroeconomic outcomes monetary policy is seeking to affect. After the experience of the 1980s, it is difficult to foresee any ready consensus on what that compact set of variables should be. Subsequent experience may provide a new basis for such a consensus, and new data to substantiate it; but that prospect remains a matter for the future, if not later still.

In the meanwhile, the Federal Reserve System has not ceased operations. Nor should it be inclined to do so, in light of the performance of both income and prices during the past half-decade. Five years of fairly steady economic growth, with inflation consistently lower than at any time since before the Vietnam War, represents no small achievement by today's standards. In the world of practical affairs, it is difficult to argue with success.

Notwithstanding the Federal Reserve's continuing formulation of money growth targets that it reports to Congress, as current law requires, and even notwithstanding the relatively high success rate in meeting the target for $M2$, it seems clear enough that the Federal Reserve System since mid-1982 has centered its monetary policy actions primarily around controlling short-term nominal interest rates.[21] In so doing, Federal Reserve decision makers have no doubt taken account of the movements of money (and perhaps credit, too); but they have also taken account of many other potential information sources, including longer-term asset prices and yields, dollar exchange rates, and numerous aspects of nonfinancial economic activity. More to the point, they have apparently proceeded in the absence of any well-articulated conceptual framework linking the interest rate as the chief policy instrument to the main macroeconomic policy objectives, or linking the associated large and diverse information base to either the policy instrument or the policy objectives. Although procedures differ in various details, the overall approach is strongly reminiscent of the practice of the 1950s and 1960s.[22]

---

[21] See Wallich (1984) for a description in different but equivalent terms.

[22] See, for example, the descriptions given by Brunner and Meltzer (1964) and Guttentag (1966).

It is therefore useful to ask why the policy approach followed at that time failed. The voluminous investigation of this question, both at the time and subsequently, supported three general conclusions: First, Federal Reserve officials systematically confused the level of interest rates as the instrument of monetary policy with the level of interest rates as an ultimate objective of monetary policy. As a result, they usually delayed too long before raising or lowering interest rate levels, and even then made changes of insufficient magnitude. Second, with no nominal quantity at the center of the policy process, the overall approach lacked an anchor to provide price stability. Although inflation was not therefore inevitable, there was little protection against it when various inflationary pressures arose. Third, once inflation did emerge, Federal Reserve officials (and many other people too) often failed to distinguish nominal from real interest rates. As a result, they often associated higher observed interest rates with a tighter policy stance even when the increase in nominal interest rates merely kept pace with, or even fell short of, rising inflationary expectations.

Are these three flaws inherent in the approach to monetary policy that the Federal Reserve System followed a quarter-century ago, and that it has apparently been following again since mid-1982? Or is it possible to design and implement monetary policy along these lines, albeit in a way that has learned from the still relatively recent past? Were the familiar failures of monetary policy under this approach in the past inevitable? Or does the experience of the last half-decade show that this kind of monetary policy can work, and work well? Research on these questions may be the best contribution economists concerned with U.S. monetary policy can now make.

■ *I am grateful to Kenneth Kuttner for research assistance; to Stanley Fischer, William Poole, Joseph Stiglitz and Timothy Taylor for helpful comments on a previous draft; and to the National Science Foundation and the Harvard Program for Financial Research for research support.*

References

Brunner, Karl, and Allan H. Meltzer, *The Federal Reserve's Attachment to the Free Reserve Concept*. U.S. Congress, House of Representatives, Committee on Banking and Currency, Subcommittee on Domestic Finance, 88th Congress, 2nd Session. Washington: U.S. G.P.O., 1964.

DeRosa, Paul, and Gary H. Stern, "Monetary Control and the Federal Funds Rate," *Journal of Monetary Economics*, April 1977, *3*, 217–231.

Eichenbaum, Martin, and Kenneth J. Singleton, "Do Equilibrium Real Business Cycle Theories Explain Postwar U.S. Business Cycles?" In Fischer, ed., *NBER Macroeconomics Annual*. Cambridge: The MIT Press, 1986.

Enzler, Jared, Lewis Johnson, and John Paulus, "Some Problems of Money Demand," *Brookings Papers on Economic Activity*, 1976, No. 1, 261–278.

Fischer, Stanley, "Contracts, Credibility, and Disinflation," mimeo, National Bureau of Economic Research, 1985.

Friedman, Benjamin M., "Money, Credit, and Interest Rates in the Business Cycle." In Gordon, ed., *The American Business Cycle: Continuity and Change*. Chicago: University of Chicago Press, 1986.

Friedman, Benjamin M., "New Directions in the Relationship Between Public and Private Debt," *Science*, April 24, 1987, *236*, 397–403.

Friedman, Benjamin M., "The Roles of Money and Credit in Macroeconomic Analysis." In Tobin, ed., *Macroeconomics Prices and Quantities: Essays in Memory of Arthur M. Okun*. Washington: The Brookings Institution, 1983.

Friedman, Benjamin M., "Time to Reexamine the Monetary Targets Framework," *New England Economic Review*, March/April 1982, 15–23.

Friedman, Milton, "Lessons from the 1979–82 Monetary Policy Experiment," *American Economic Review*, May 1984, *74*, 397–400.

Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, 1976, No. 3, 683–730.

Goldfeld, Stephen M., "The Demand for Money Revisited," *Brookings Papers on Economic Activity*, 1973, No. 3, 577–638.

Goldfeld, Stephen M., and Daniel E. Sichel, "The Demand for Money." In Friedman and Hahn, eds., *Handbook of Monetary Economics*. Amsterdam: North-Holland Publishing Company, forthcoming.

Guttentag, Jack M., "The Strategy of Open Market Operations," *Quarterly Journal of Economics*, February 1966, *80*, 1–30.

Judd, John J., and John L. Scadding, "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature*, September 1982, *20*, 993–1023.

Lombra, Raymond, and Michael Moran, "Policy Advice and Policymaking at the Federal Reserve." In Brunner and Meltzer, eds., *Monetary Institutions and the Policy Process*. Amsterdam: North-Holland Publishing Company, 1980.

Okun, Arthur M., "Efficient Disinflationary Rules," *American Economic Review*, May 1978, *68*, 348–352.

Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, *84*, 197–216.

Porter, Richard D., Thomas D. Simpson, and Eileen Mauskopf, "Financial Innovation and the Monetary Aggregates," *Brookings Papers on Economic Activity*, 1979, No. 1, 213–237.

Roley, V. Vance, "Money Demand Predictability," *Journal of Money, Credit and Banking*, November 1985, *17*, Part 2, 611–641.

Sachs, Jeffrey D., "The Dollar and the Policy Mix: 1985," *Brookings Papers on Economic Activity*, 1985, No. 1, 117–185.

Scholl, Russell B., "The International Investment Position of the United States in 1986," *Survey of Current Business*, June 1987, *67*, 38–45.

Simpson, Thomas D., and Richard D. Porter, "Some Issues Involving the Definition and Interpretation of the Monetary Aggregates." In *Controlling Monetary Aggregates III*. Boston: Federal Reserve Bank of Boston, 1980.

Solomon, Robert, "The United States as a Debtor in the Nineteenth Century," mimeo, The Brookings Institution, 1986.

Stock, James H., and Mark W. Watson, "Interpreting the Evidence on Money-Income Causality," mimeo, National Bureau of Economic Research, 1987.

Wallich, Henry C., "Recent Techniques of Monetary Policy," Federal Reserve Bank of Kansas City, *Economic Review*, May 1984, 21–30.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 73–100

Monetary Policy Lessons of Recent Inflation and Disinflation

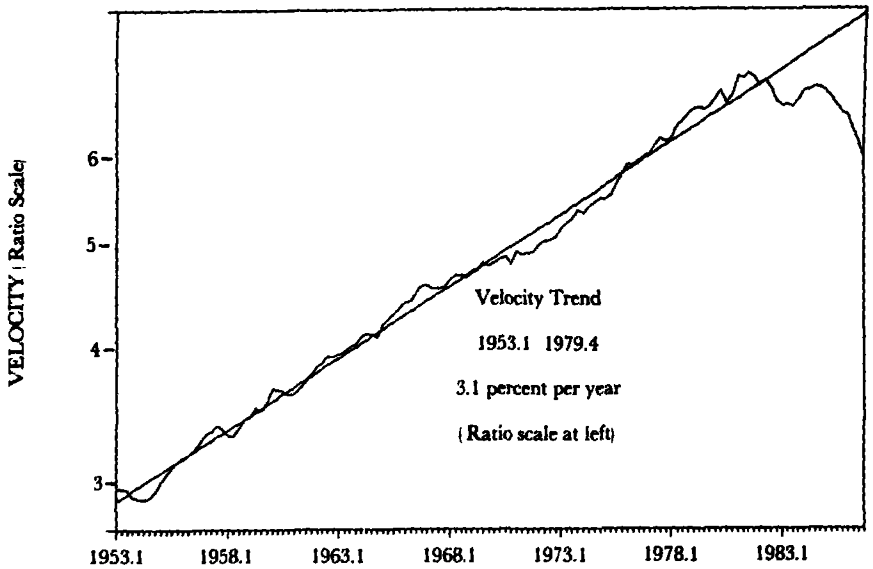
William Poole

By 1975, most economists agreed that the money demand function in the United States was reasonably stable and could serve as a reliable basis for the formulation of monetary policy. Those suspicious of monetarism were on the defensive in the light of the apparently inexorable increase of $M1$ velocity of about three percent per year with deviations of only a few tenths of a percent. Experience with rising money growth and rising inflation through 1980 only confirmed monetarist views.

As Figure 1 shows, all this had changed by 1986. With disinflation in the 1980s, $M1$ velocity departed convincingly from its 1953–79 trend. The money demand function seems to have fallen apart and is apparently not a reliable basis for monetary policy after all. Consequently, the first major section following this introduction is devoted to the decline of velocity after 1981 and the demise of the "standard" money demand function. Some authors, including (I believe) Benjamin Friedman, conclude that this evidence leaves us so uncertain about the nature of money demand that policymakers cannot rely at all on any presumed money demand regularities. In contrast, my working hypothesis is that the money demand continues to be a stable function of relatively few variables, but that the interest elasticity of money demand is substantially higher than previously thought. (Others have sought money demand explanations in terms of regime shifts, but that literature will not be discussed here.)

Events have not been kind to Keynesian monetary policy positions either. Keynesians tend to concentrate on interest rates—especially real interest rates—as the best guide to the effects of monetary policy on the economy. The real rate is discussed

■ *William Poole is Professor of Economics, Brown University, Providence, Rhode Island.*



Data are from the Data Resources, Inc. data bank unless otherwise indicated.
Velocity is the ratio of current dollar GNP to M1.
GNP
1953–1986: DRI Series GNP
M1
1953–1958: Friedman–Schwartz series spliced by author to series for 1959–1986.
1959–1986: DRI Series MNY1

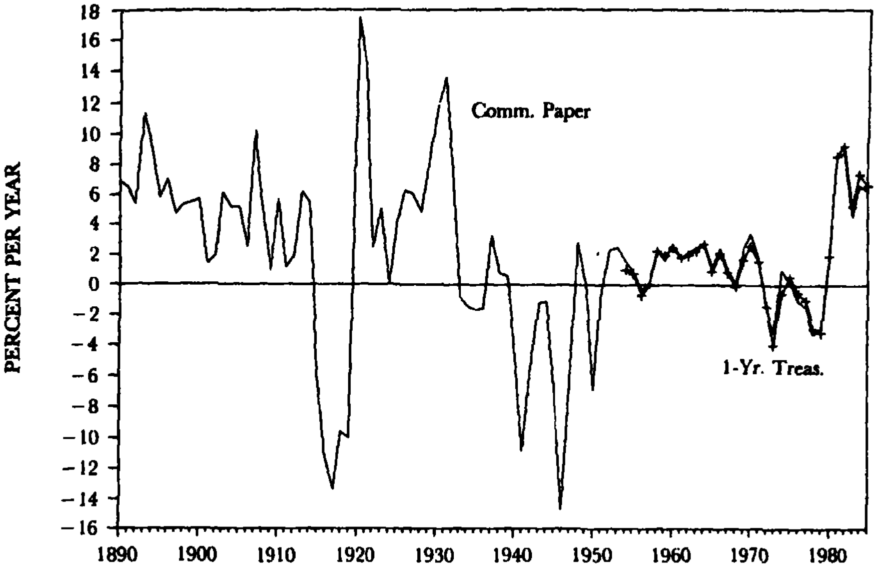
Fig. 1. Velocity and velocity trend, quarterly, 1953–1986

in the third section of the paper. As can be seen in Figure 2, real rates rose sharply i late 1979 and remained high through 1985.[1] The severity of the 1981–82 recession (measured, say, by the peak unemployment rate of almost 11 percent) is consiste qualitatively with the high real rate of interest, but the vigorous growth of output a employment in 1983–84 are not.

The fourth section is devoted to an analysis of the importance of mar expectations about monetary policy, with a particular focus on how market expec tions were influenced by the Federal Reserve's introduction of new monetary cont procedures in October 1979. The fifth section of the paper contains an evaluatio

---

[1] Each observation of the real rate in this figure is the annual average short-term interest rate for year t the percentage change in the annual average of the Consumer Price Index from year t to year $t + 1$. figure shows the real rate calculated from the commercial paper rate for the entire period and fro one-year U.S. Treasury yield for 1954–85. Ideally, to match the inflation rate measured by annual av of the CPI the interest rate should be the yield on a one-year security, but the only series readily ava back to 1890 is the commercial paper rate. The fact that there is so little daylight between the two rea series for 1954–86 suggests that it is unlikely that the basic message of Figure 2 for 1890–1953 is affected by using the commercial paper rate.



Data are from the Data Resources, Inc. data bank unless otherwise indicated.
Nominal *Prime Commercial Paper Rate*
1890–1960: *Historical Statistics of the United States*, Series X-445 (4–6 mo.)
1961–1985: DRI, Series RMCML6NS (6 mo.)
Nominal *Yield on Treas Securities at Constant Maturity of 1 Year*
1954–1985: DRI Series RMGFCM1NS
Real yields obtained by subtracting percentage change in *Consumer Price Index, All Items*
1890–1912: *Hist Stat* Series E-135
1913–1986: DRI Series CPIU

Fig. 2. One year rate of interest, ex post

the contribution of monetary policy to the disinflation of the 1980s. In the sixth and final section I conclude with some comments on the implications of the previous sections for the debate on monetary rules.

The Decline of Velocity After 1981

At least for some monetarists, and I believe for most Keynesians also, the behavior of velocity after 1981 was a surprise, and there is much to be learned, or relearned, from understanding why. Because the majority of economists have traditionally favored *M*1 and because the issues I discuss are not peculiar to *M*1, I will confine my analysis to *M*1 and *M*1 velocity, and not take up the issue of whether definitions of money other than *M*1 should be employed.

The Standard Money Demand Function

The money demand function was the subject of an enormous amount of work in the 1960s and 1970s. Most economists settled on a formulation that made real money demand M/P a log linear function of real income y and one or more interest rates R, as in equation 1 below. For estimation purposes the function was often stated in velocity form, as in equation 2.

$$(1) \qquad \log(M/P) = -b_0 + b_1\log(y) - b_2\log(R)$$

$$(2) \qquad \log(V) = \log(yP/M) = b_0 + (1 - b_1)\log(y) + b_2\log(R)$$

Most quarterly specifications included a lagged adjustment mechanism of some sort, with the Koyck form most often employed.[2] There was general, though not unanimous, agreement that $M1$ was the most appropriate monetary aggregate and that one or more short-term interest rates worked better than a long-term interest rate. It appeared that the interest elasticity b_2 was relatively low (in absolute value), perhaps in the neighborhood of 0.15 to 0.25. It also appeared that the real income elasticity b_1 was somewhat below unity, perhaps in the neighborhood of 0.6 to 0.8 (Goldfeld, 1973).

Confidence in the stability of the money demand function began to erode in 1975 as velocity rose more rapidly than could be explained by the standard money demand function. The problem was "The Case of the Missing Money," which was the title of a paper by Goldfeld (1976). I believed at the time, and still believe, that estimates of money demand parameters were much less precise than estimated standard errors indicated.[3] The problem is that an estimation period with an initial year after 1946—the initial year for many estimations is 1952 or 1953—is dominated by upward drifts in all the variables that enter the money demand function, as shown in Table 1 which reports a correlation matrix for the relevant variables for 1947:$Q1$ to 1986:$Q4$.[4] The unreliability of money demand estimates, due largely to multicollinearity, is indicated by the fact that widely differing constraints on the income elasticity yield very different interest elasticities but only small differences in the regression R^2 or standard error of estimate.[5] This point was discussed in the 1970s, but not enough attention was paid to it.

---

[2] In the Koyck form, lagged adjustment is captured by entering a lagged dependent variable in a regression. A less restrictive model employs one or more lagged independent variables, but using this more general distributed lag model rapidly uses up degrees of freedom.

[3] My argument was spelled out in Poole (1970). The same point is made much more forcefully and carefully by Cooley and LeRoy (1981). They argue (p. 836), "[t]he data are such that a modestly energetic specification search will give back almost whatever interest elasticity one wishes to extract..." I accept this argument if the empirical analysis is confined to U.S. data for the postwar period. However, the argument surely goes too far if the data universe includes, for example, a number of hyperinflation cases or a cross section of countries with widely differing inflation rates.

[4] The first quarter for which quarterly GNP data are available is 1947:$Q1$. The correlation matrix for 1953:$Q1$ to 1980:$Q4$ is essentially identical to Table 1.

[5] In Poole (1970) I showed that the R^2 of the money demand function was little affected by varying an income elasticity constraint between 1.0 and 3.0. However, the estimated interest elasticity is highly sensitive to the income elasticity constraint. Using data since 1970 does not change these results in any significant fashion.

Table 1
Correlation matrix, major macro variables quarterly, 1947.1–1986.4

	Time	GNP	GNP82	*P*GNP	*M*1	*M*1/*P*	Aaa	*TB*	*VEL*
Time	1.000	0.993	0.993	0.967	0.970	0.400	0.973	0.911	0.987
GNP	0.993	1.000	0.979	0.989	0.989	0.387	0.970	0.893	0.975
GNP82	0.993	0.979	1.000	0.939	0.945	0.427	0.964	0.928	0.991
*P*GNP	0.967	0.989	0.939	1.000	0.996	0.348	0.949	0.846	0.939
*M*1	0.970	0.989	0.945	0.996	1.000	0.431	0.941	0.837	0.933
*M*1/*P*	0.400	0.387	0.427	0.348	0.431	1.000	0.290	0.233	0.302
Aaa	0.973	0.970	0.964	0.949	0.941	0.290	1.000	0.932	0.973
TB	0.911	0.893	0.928	0.846	0.837	0.233	0.932	1.000	0.943
VEL	0.987	0.975	0.991	0.939	0.933	0.302	0.973	0.943	1.000

All data in logarithmic form except for time dummy

Time: Time dummy
GNP: Gross national product, current dollars
GNP82: Gross national product, 1982 dollars
*P*GNP: Gross national product implicit price deflator
*M*1: *M*1 money stock
*M*1/*P*: Ratio of *M*1 to *P*GNP
Aaa: Aaa bond yield
TB: 13-week treasury bill rate, secondary market
VEL: *M*1 velocity = ratio of GNP to *M*1

Understanding Velocity Trends

Analysis of velocity trends has a long history, and many examples could be cited. Irving Fisher (1963, pp. 79–88) argued that improvements in communications and transportation would raise velocity. Milton Friedman and Anna J. Schwartz (1963, pp. 639–45) argued that velocity had a secular downward drift through World War II due to an income elasticity of demand for money above unity, and they attributed the postwar rise in velocity to increased confidence concerning economic stability.[6] These hypotheses were tested by Gould and Nelson (1974) who used time series analysis to show that velocity over the period from 1867 to 1970 had statistical properties closely resembling a pure random walk without drift.

Recent review articles by Laidler (1980) and Judd and Scadding (1982) have reflected the thrust of mainstream money demand research, which has been to find economic explanations for changes in velocity. That search is exactly the correct enterprise. Analyzing time series properties of velocity is not a substitute for economic explanations, but instead provides a statistical base with which economic explanations must be consistent.

---

[6] Friedman and Schwartz concentrated on *M*2 velocity, whereas in this paper I am confining my attention to *M*1 velocity. Although these two velocities have different levels and different secular trends, changes in their growth rates from one subperiod to another are quite similar.

Table 2
Velocity regressions, first differences of annual data

Period	*Constant* *(T-Stat)*	*Coefficient* *(T-Stat)*	*Adj.* R^2	*Std. Err.*	*DW*
1916–86	0.007 (1.050)	0.238 (3.005)	0.103	0.056	1.44
1919–46	−0.030 (−2.001)	−0.161 (−0.687)	−0.020	0.071	1.81
1947–81	0.032 (6.723)	0.071 (1.350)	0.024	0.024	1.90
1916–57	0.002 (0.182)	0.284 (1.774)	0.050	0.070	1.49

Dependent variable: first difference log *M1 velocity*
Independent variable: first difference log Aaa *bond yield*

Developments in econometrics theory in the 1970s suggested that it is dangerous to estimate equations using data subject to significant trends. Variables that are in fact unrelated to each other can display spurious correlations when they are subject to independent trends. Thus, authors increasingly worked with detrended data or ran regressions with data in first difference form. For a postwar sample period, using differenced data in a regression with velocity as the dependent variable invariably yields an estimated equation with a statistically and economically significant constant term and a very low interest elasticity. The constant term in the regression picks up the average velocity increase of about 3 percent per year and interest rates add relatively little explanatory power.

Although economists generally hesitate to put a time dummy into a regression as an explanatory variable for fear that it will pick up explanatory power rightfully due to one or more economic variables, few seem to worry about leaving a constant term in a first difference regression. This practice is peculiar, in that a constant term in a regression using the first differences of the data is, of course, equivalent to a linear time trend specification in a regression using the levels of the data. The issue is critical for estimating a money demand function because of the need to determine how much of the postwar increase in velocity should be attributed to rising interest rates and how much to some other economic variable or to a time dummy.

Table 2 reports very simple *M*1 velocity regressions using first differences of natural logs of annual data 1916–86. For reasons to be discussed later, the interest rate variable in the regressions is Moody's Aaa long-term bond yield and the real income elasticity has been constrained to 1.0 by not including real GNP as a regressor. The sample periods chosen for Table 2 reflect the facts that: a) the end points of the available annual *M*1 data are 1915 and 1986; b) the peak years for velocity before and after World War II were 1918 and 1981; c) the trough year for velocity was 1946; and d) velocity was almost identical in 1915 and 1957. Thus, the regressions cover periods of falling velocity, rising velocity, and unchanged velocity.

Table 2 tells a straightforward story. When velocity is falling, as from 1919 to 1946, the constant term picks up the decline. In fact, for this period the interest rate coefficient has the wrong sign. When velocity is rising, as from 1946 to 1986, the constant term picks up the increase. The interest rate coefficient has the correct sign, but its t-statistic falls short of conventional levels of statistical significance. When the overall velocity drift is zero, as for 1916–57, the constant term is very small and the coefficient on the interest rate is considerably higher. The regression for the full period 1916–86, which has a small velocity increase, has a higher constant term and lower interest rate coefficient than does the regression for 1916–57. I conclude that the first difference specification with a constant term is defective; parameter estimates are very sensitive to the sample period and the adjusted R^2 is never very high.

The 1970s research on money demand, taken with other developments, affected policy advice late in the decade. The research findings, to summarize, included instability of structural models of money demand, low estimates of the interest elasticity, and the apparently highly predictable velocity increase of about 3 percent per year. The other developments included the growing use of time series (ARIMA) forecasting models and the increasing urgency of controlling inflation. A number of economists, myself included, began to base monetary policy advice on two assumptions: that velocity would continue to grow at a 3 percent rate and that the interest elasticity was essentially zero.

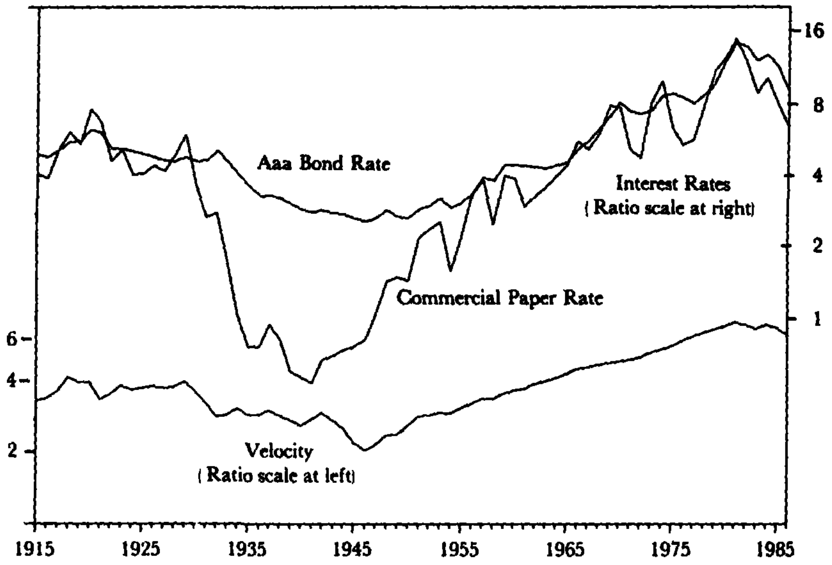
We committed an avoidable error: we insisted on economic explanations for velocity trends, but in analyzing monetary policy we relied on a velocity time trend. The standard money demand function foundered in the mid- and late 1970s and the reliable velocity trend broke down in the early 1980s. Where does all this leave us?

Repairing the Money Demand Function

The first thing to say is that it is hopeless to attempt to resolve uncertainties about money demand through analysis of postwar U.S. data alone; the multicollinearity problem is overwhelming. Second, transforming the data to first difference form yields, I believe, a serious underestimate of the interest elasticity of demand. Let me now explore this second point.

Figure 3 shows $M1$ velocity, the commercial paper rate, and the Aaa bond yield for annual data 1915–86 plotted on a ratio scale. Velocity is an extremely smooth series compared to the two interest rate series. There has been a long debate over which interest rate to use in the money demand function, but on the face of it the long rate would seem to have the edge. The short rate simply looks too volatile.

The lagged adjustment model is the usual way to reconcile the relatively low volatility of velocity with the relatively high volatility of the short interest rate. Goldfeld (1973) estimated an adjustment speed of 25–30 percent per quarter. However, with annual data the conventional adjustment lag argument is not convincing. Why should economic agents hold idle money balances when non-monetary assets bearing a higher interest rate are so readily available?



Data are from the Data Resources, Inc. data bank unless otherwise indicated.

Commercial Paper Rate

Same as Figure 2

Aaa Bond Rate

1915–18: Macauley Railroad bond yield, *Hist Stat* Series X-476, +0.22 to splice this series to the next one.
1919–46: Moody's Aaa Corp Bond Yield, *Hist Stat* Series X-477
1947–86: Moody's Aaa Corp Bond Yield, DRI Series RMMBCAAANS

Velocity is ratio of current dollar GNP to M1.

GNP in Current Dollars

1915–1928: *Hist Stat* Series F-1
1929–1986: DRI Series GNP

M1

1915–1958: Friedman–Schwartz series spliced by author to series for 1959–1986.
1959–1986: DRI Series MNY1

Fig. 3. Velocity and interest rates, annually, 1915–1986

The Case for the Long Rate

Table 3 reports annual velocity regressions in which the income elasticity is constrained to be 1.0 (for reasons discussed later). These regressions yield a much higher estimate of the interest elasticity for the long rate than for the short rate. Many others have found this same result but, strangely, no one seems to have been puzzled by it. Under the expectations theory of the term structure of interest rates, the long rate is a weighted average of expected future short rates, and consequently the full-adjustment (or long-run) interest elasticity with respect to the long rate ought to equal that with respect to the short rate.

To understand this point, suppose an economy operates for a long time with zero inflation and short and long rates in the neighborhood of 3 percent, and then adjusts

Table 3

Velocity regressions, levels of annual data

	Coefficients (T-Stat)					
Period	*Constant*	*RAAA*	*RPAP*	*Adj.* R^2	*Std. Err.*	*DW*
1915–86	0.227	0.669		0.908	0.098	0.45
	(6.443)	(26.542)				
	0.959		0.316	0.732	0.168	0.22
	(29.132)		(13.594)			
Beg *R*		4.84%	4.01%			
End *R*		9.02%	6.39%			
1915–64	0.365	0.584		0.702	0.099	0.47
	(4.956)	(10.796)				
	0.998		0.188	0.648	0.108	0.36
	(45.842)		(9.556)			
Beg *R*		4.84%	4.01%			
End *R*		4.40%	3.96%			
1920–68	0.299	0.645		0.718	0.108	0.40
	(4.776)	(11.103)				
	0.998		0.208	0.653	0.119	0.30
	(41.492)		(9.568)			
Beg *R*		6.12%	7.50%			
End *R*		6.18%	5.90%			
1921–86	0.271	0.670		0.914	0.099	0.42
	(6.435)	(26.237)				
	0.963		0.326	0.757	0.167	0.23
	(29.435)		(14.278)			
Beg *R*		5.97%	6.62%			
End *R*		9.02%	6.39%			

Dependent variable: log *M1 velocity*
Independent variables: RAAA = log Aaa *bond yield*
RPAP = *log commercial paper rate*
Beg R = *Interest rate in first year of period*
End R = *Interest rate in last year of period*

to a long-lasting inflationary equilibrium with an inflation rate of 10 percent and short and long rates in the neighborhood of 13 percent. Using these hypothetical data, estimates of the interest elasticity from short rates and from long rates should be identical because both rates have risen by the same amount.

Now consider an actual estimation. If the sample period is long, and if the general level of rates changes significantly over the sample, the correlation of short rates and long rates will be very high. For annual U.S. data 1915–86 the correlation between the logs of the two rates in Figure 3 is 0.868. The estimated interest elasticity of money demand ought not to depend very much on which interest rate is used in a regression.

In fact, in the regressions in Table 3 the long rate not only yields an estimated elasticity twice the elasticity of the short rate but also a higher R^2. I conclude that the long rate rather than the short rate belongs in the money demand function. The short rate yields a lower interest elasticity because the short rate measures the true opportunity cost variable with error.[7]

It should be emphasized that this errors-in-variables argument could go either way. The fact that the short rate has a higher variance than the long rate does not automatically mean that the short rate will have a lower regression coefficient.[8] If velocity were in fact a function of the short rate, then velocity would be much more volatile than it is. In this case, the long rate could yield a lower regression coefficient than the short rate because the long rate would measure the short rate with error.

The argument rationalizing the better fit of the long rate is that changes in the amount of money held relative to income require that economic agents make capital investments of various kinds in order to economize on money. For example, to reduce cash balances a corporation may have to invest in personnel and computers. Banks may open new offices to reduce travel time ("shoe-leather costs") for their customers. These investments will not be undertaken unless an increase in interest rates is judged to be permament—long-lasting enough relative to the life of the investment. Nor will a decline in interest rates judged to be transitory lead agents to abandon past investments that economize on the holding of money. The long-term interest rate reflects expectations about future short-term interest rates, and so measures the appropriate cost variable in the money demand function more accurately than does the short-term interest rate. Although the long rate, on this argument, is the appropriate cost variable for the money demand function, in managing money balances agents may well be switching between money and short assets.[9]

Money as a Buffer

The usual conception of money demand is that the error term in a money demand function is statistically independent of other variables and disturbances in the system of equations describing the economy. Such a disturbance will cause changes in one or more endogenous variables in the system. A money demand disturbance will have the same effect on the system as would a shift of the same size in the constant term of the money demand equation.

The buffer stock view of money demand has very different implications. In the short run, money is a "slack" variable that absorbs disturbances originating elsewhere. Part of an agent's money balance at any given time reflects what Keynes called the

---

[7] More precisely, the short rate measures the true opportunity cost with greater error than does the long rate. Goodfriend (1985) provides an analytical argument that estimated lagged adjustment terms in the standard money demand function are a spurious result of measurement error and that measurement error leads to biased estimates of elasticities. He does not, however, provide elasticity estimates corrected for the problems he notes.

[8] The standard deviations of the log levels of the commercial paper rate and AAA bond rate for 1915–86 are 0.46 and 0.88, respectively.

[9] This paragraph should not be read as claiming that real assets are irrelevant to money demand. There is ample evidence that agents flee from both money and fixed income assets into goods during periods of hyperinflation. During moderate inflations, as in the United States in the 1970s, agents may economize on money by holding inventories of goods or accelerating purchases of consumer durables.

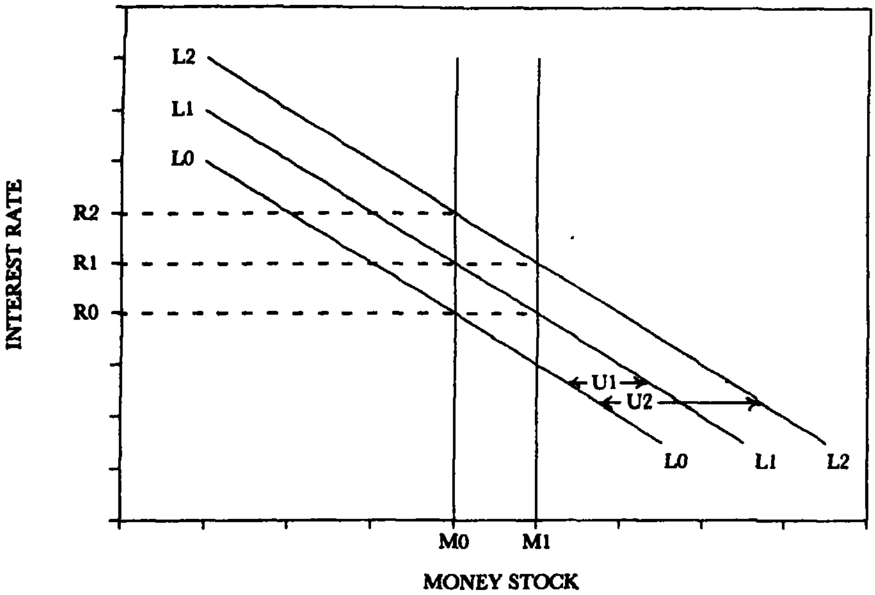


Fig. 4. Disturbances to money demand and supply

"precautionary motive" for holding money. Another part reflects a particular realization of net cash flows from other activities which leave an agent's money balance transitorily high or low. Milton Friedman's phrase "a temporary abode of purchasing power" carries the right flavor. On a formal level, the stochastic inventory model by Miller and Orr (1966) rings true to me. Laidler (1985) provides a short discussion of the buffer stock argument and a bibliography of relevant papers.

The implications of the buffer stock model can be seen by examining several examples with the aid of Figure 4. The initial money demand function is L_0, and we assume that the money stock jumps unexpectedly from M_0 to M_1 due to an open market operation by the central bank. The private sector holds more money and fewer bonds than before. However, in the absence of any other disturbance there is no reason for anyone to do anything in the short run. Agents may not know the source of the extra money and have no reason to believe that the disturbance is permament. Consequently, in the short run the interest rate remains at R_0 and the flow of spending does not change. In the language of Keynes' speculative demand for money, changes in money supply that do not alter interest rate and income expectations are willingly held with little or no effect on interest rates and spending.

The econometrician, however, interprets this case as involving a disturbance U_1 to the money demand function shifting it from L_0 to L_1 and assumes that U_1 is independent of the increase in the money stock. That interpretation is the only conventional way to reconcile the observation of a money stock of M_1 with the interest rate of R_0. Because the money demand disturbance is interpreted as independent, the analyst may conclude that it is very fortunate that the central bank accommodated the disturbance; if it had not done so, then the interest rate would have risen to R_1.

For another example, consider a case in which new information—such as election results, a policy announcement by the government, or a statistical release reporting last month's increase in the money stock—leads the market to bid up interest rates. The initial equilibrium has money demand function L_0, money stock M_0, and interest rate R_0. The interest rate rises instantaneously to R_1 when the market receives new information, but at this instant the money stock does not change. On the conventional view there has been a money demand disturbance U_1 equal to the horizontal distance between money demand functions L_0 and L_1 and this disturbance has caused the increase in the interest rate to R_1.

Surely, however, if the central bank had coincidentally increased the money stock by the size of this disturbance—that is, from M_0 to M_1—the interest rate would have gone up anyway to clear the credit markets in the wake of receipt of new information. If the coincidental increase in the money stock had actually occurred, the conventional view would be that a money demand disturbance U_2 had shifted the money demand function from L_0 to L_2. That is the only possible conventional interpretation to rationalize an interest rate of R_1 and a money stock of M_1. An observer holding to the conventional view of independent money demand disturbances will believe that it is highly desirable that the central bank followed an accommodative policy permitting the money stock to rise to M_1, for otherwise the disturbance (which is really a credit market disturbance) shifting the money demand function to L_2 would have pushed the interest rate up to R_2.[10]

This model leaves the econometrician in a difficult situation. In the examples discussed above, money demand disturbances are correlated with variables and/or disturbances in the credit market and with money supply disturbances. There is no obvious way to estimate the aggregate money demand function to get around the problem that money demand disturbances are correlated with other variables and disturbances in the system. Moreover, because expectations (rational or otherwise) affect interest rates all information relevant to expectations is likely to be correlated with the money demand disturbances identified by the econometrician. Finally, the adjustment process through which departures of money balances from long-run equilibrium levels have effects on the endogenous variables in the money demand function is likely to be highly variable.

These problems seem likely to be especially serious when estimating regressions in first difference form. A substantial part of the variance of quarter-to-quarter and year-to-year changes in interest rates reflects disturbances elsewhere in the economy that are correlated with (conventionally defined) money demand disturbances. An example was given above where new information bids up interest rates with no immediate change in the money stock. The result is a substantial bias toward zero in the estimated interest elasticity of money demand in a first difference equation.

---

[10] The only way, to my knowledge, to represent this analysis in a formal system of equations is to assume that the disturbance in the money demand equation is correlated with variables and/or disturbances in other equations. In Poole (1982) I explored these ideas in a formal but incomplete fashion.

Although transitory changes in the money stock are willingly held in the short run, over the longer run the economy does adjust to sustained changes in the money stock. These longer run adjustments are likely to be better estimated in a regression using the levels of the data than in one using the first differences because the noise in the variables will be smaller compared to their systematic changes. The smaller the noise in the variables, the smaller will be the bias in estimates of regression coefficients.

In a levels regression, however, the residuals are highly serially correlated creating two problems for estimating money demand. The first is that the estimates of the standard errors are biased downward. That is not too serious an issue because under usual assumptions the coefficient estimates themselves are not biased. However, if the residuals are in fact generated by a random walk, the levels regression may yield biased coefficient estimates. With a random walk, even one without drift, the mean of the levels residuals is unlikely to be zero and may be time-dependent.[11]

Suppose the true money demand residuals starting with the first observation (1915) accumulate so that the last observation (1986) has a residual that is positive and relatively large. The true mean of the residuals is positive, but the estimated regression will have coefficients such that the estimated residuals have a zero mean. Thus, in a levels regression the realized drift in the residuals will be assigned to any variable in the regression that has a drift (up or down) over the sample period.

Consider a levels regression with velocity as the independent variable. If the residuals are correctly modeled as a random walk, and if they have an upward (downward) drift over 1915–86, then the interest rate will have an upward (downward) biased coefficient because the rate rose over this sample period. One way to avoid this bias is to use a sample period for which the interest rate is about the same at both ends. Thus, in Table 3 the interest elasticities estimated from the Aaa bond rate over 1915–64 and 1920–68 are least likely to be biased. From these results I conclude tentatively that the interest elasticity of demand for real money balances is about 0.6 in absolute value.

Although there is some protection against bias in the estimate of the interest elasticity from using sample periods in which there was no overall increase or decrease in interest rates, that protection is not available when trying to estimate the income elasticity because income almost always rises. In working on this paper, I made no effort to estimate the income elasticity of money demand because I am not convinced that I could learn anything from the attempt. It seems likely that convincing evidence on the income elasticity will have to come from studies using micro data and country cross section data.[12]

---

[11] One of several plausible interpretations of the economics behind a random walk in the residuals is that of random technical change in payments mechanisms and practices. It is not illogical to suppose that transactions costs follow an random walk due to random technical developments in data processing and so on. Because these costs are not modeled explicitly they show up in the error term of the money demand function.

Other Money Demand Issues

In the quest for a stable money demand function it remains an open question as to whether the best definition of money in the United States is $M1$, $M2$, or some other M. Various definitions of the money stock are highly correlated with each other over long periods, and it seems probable that there is no way to extract from time series data the answer as to which one is best. This issue deserves study, but it seems much less important than the interest elasticity and velocity trend issues discussed above.

An even more difficult issue is that of the short-run adjustment mechanism. The approach underlying the standard money demand function relies on a model of adjustment by an individual agent who brings actual balances into line with desired balances. The problem is that this model, whether in Koyck or distributed lag form, does not invert properly to the case in which exogenous changes in the money stock yield lagged adjustments in the arguments of the money demand function—in the price level, nominal interest rates, and real GNP. Because most central banks follow a policy of monetary accommodation in the short run, some of the data are generated by a process in which the aggregate money stock adjusts passively to changes in the arguments. However, the data also include observations in which the money stock changes actively and prices, incomes, and interest rates adjust over time. This mixture of observations involving endogenous and exogenous changes in the money stock creates vexing econometric difficulties.

There is no obvious way to extract from the data these different types of adjustments. Here, even micro data may be of little assistance because macro lags may arise from the accumulation of relatively short lags at the micro level. Moreover, money demand adjustment mechanisms depend importantly on how expectations are related to observable data and so are almost certain to be unstable for the reasons detailed by Lucas (1976) that have come to be called the "the Lucas critique." The Lucas argument applies to money demand because agents will treat monetary disturbances as transitory if the central bank systematically reverses them; but agents will treat such disturbances as permanent if the central bank does not reverse them. Thus, agents' short-run responses to changes in the money stock will not be invariant to the policy chosen; and, as Lucas emphasized, estimates of lagged responses under one policy cannot provide useful information on responses under a different policy.

To summarize these thoughts about money demand, a typical money demand function of the late 1970s with a low interest elasticity and a time trend has performed miserably in the 1980s. However, experience in the 1980s fits a model with no time trend and a considerably higher interest elasticity than was generally accepted a few years ago. Finally, no one has constructed a convincing model of lagged adjustment.

---

[12] Bordo and Jonung (1986) provide cross section evidence for the hypothesis that the income elasticity of demand for real money balances is above unity for relatively low per capita income countries and below unity for relatively high per capita income countries. However, these authors do not report numerical estimates of income elasticities.

and economists must conclude that we know essentially nothing about adjustment speeds—except that they are faster during hyperinflations than at other times—and nothing about the stability or regularity of adjustment mechanisms. Efforts to estimate long-run elasticities are much more likely to be successful than those to estimate adjustment speeds.

The Real Rate of Interest

The behavior of the real rate of interest in the 1980s has been every bit as much a surprise as has the behavior of velocity. Fama (1975) argued that the expected real rate is a constant. Few economists accepted that hypothesis as literally true, but the evidence did support the proposition that most of the variance of nominal short-term interest rates reflected variance in inflationary expectations rather than variance in the expected real rate of interest (Fama and Gibbons, 1982). Depending on the choice of short-term rate, the real rate for 1953–73 was 1 to 2 percent and for 1973–78 was 0 to −2 percent. For 1980 to 1985, however, the real rate—both expected and realized—was in the 4 to 8 percentage range.[13] These rough figures are satisfactory for the purpose at hand; everyone agrees the real rates were dramatically higher in the early 1980s than they were during the previous several decades.[14]

In 1981–82 it was logical to attribute high real rates to monetary stringency, just as it was logical to attribute abnormally low real rates in the late 1970s to monetary laxity. On this interpretation, no new lesson for monetary policy arose from either experience. The period of accelerating inflation and abnormally low real rates in the late 1970s would not have occurred if the Federal Reserve had held money growth to, say, 4 percent per year. The deceleration of money growth and high real rates in 1981–82 would not have occurred if the inflationary monetary policy of the late 1970s had not occurred. Or taking the initial conditions of 1980 as given, a more gradualist monetary policy in 1981–82 would presumably have yielded a smaller increase in real rates and a more shallow recession than the one actually experienced.

Real Determinants of the Real Rate

Keynesians tend to look toward the real rate of interest as the best measure of the stance of monetary policy, but that approach is suspect if real rates are importantly influenced by nonmonetary events. Whatever may be the validity of the monetary explanation for high real rates in 1981–82, that explanation certainly did not fit the conditions following the recession trough in late 1982. Most Keynesians predicted a slow recovery from the recession because real rates of interest were higher in the early 1980s than at any time since the onset of the Great Depression. In the past high real rates have almost invariably been associated with recession and falling prices, as in 1929–32, 1920–21, 1907, and 1893–94. (See Figure 2.) The combination of high real rates and vigorous cyclical expansion during 1983–84 and continuing, though slow, expansion through 1985–87 is very unusual in the light of U.S. experience over the past century.

---

[13] The real rate shown in Figure 2 is ex post whereas the ex ante real rate is the relevant concept for most behavioral relations in economics. However, market data on inflationary expectations do not exist and survey data are suspect; thus, ex ante real rates cannot be measured accurately. But ex ante rates must have risen substantially in the early 1980s; assuming that they did not requires acceptance of the hypothesis of a prolonged string of large inflation forecast errors in the same direction. Such a string of forecast errors must seem implausible even for those skeptical of rational expectations.

[14] For real rate estimates and charts see Huizinga and Mishkin (1986) and the discussion of the Huizinga-Mishkin paper by Schwert (1986).

The high real interest rates during 1983–85 almost certainly reflected one or more real disturbances, and the case for a real explanation is reinforced by the behavior of the real exchange rate. Based on the multilateral real exchange rate series calculated by Feldstein and Bacchetta (1987), the dollar appreciated by over 60 percent from its 1980 average to its monthly high in February 1985. For 1983–85, the behavior of the real exchange rate, of the real economy, and of the real interest rate are simply not consistent with a monetary explanation. Indeed, the monetary explanation for 1981–82 may have to be supplemented given that the real rate of interest was abnormally high by the standards of previous postwar recessions, although not by the standard of 1929–32.

What were the real disturbances? I favor the view that the primary real disturbance was the 1981 change in U.S. tax law which had the effect of substantially increasing the real after-tax rate of return on new business investment. Others hold to the view that the substantial federal budget deficit that began in 1981 and continues to this writing is responsible. Campbell and Clarida (1987, pp. 30–37) provide a summary of these positions, as well as a bibliography of many of the contributions to this debate. For the purpose at hand, the source of the real disturbance is irrelevant. There was a real disturbance, and it was large. Under these conditions it is clear that the real rate of interest cannot serve as a measure of the stance of monetary policy.

Another much-discussed class of real disturbances is the "supply shock," and there is a substantial literature on energy, food, raw materials, exchange rate, productivity, and other shocks of the 1970s. Wilcox (1983) provides an analysis and bibliography for the effects of supply shocks, describing how they reduce output and the real rate of interest, and raise the price level. However, a monetary hypothesis competes with the supply-shock hypothesis for explaining the 1970s because the alleged supply shocks occurred at times when the economy was stressed by accelerating money growth. Monetary policy was clearly inflationary in 1972–73 before the mideast war and OPEC oil shock in October 1973, and again in 1977–78 before the overthrow of the Shah of Iran and the second oil shock of 1979. In both episodes inflation was unambiguously rising *before* the oil shocks.

My own position is that real disturbances are much overrated as a source of the problems of the 1970s, and that the real effects of inflationary monetary policy are much underrated. On this view, the supply shocks would have had relatively modest effects if they had occurred in an economy characterized by both general price stability and confidence in that stability. Inflationary monetary policy made the economy vulnerable to bad news; the shocks surely were themselves a problem but their main significance was to trigger the inflationary bomb that was bound to go off anyway.

Monetary Policy Implications of Real Disturbances

The issue of the appropriate monetary policy response to a real disturbance, either through a discretionary policy adjustment or through a monetary rule which has a response built in, is too large to be discussed satisfactorily here. But the dimensions of the problem are larger than I had previously thought. My estimate of the magnitude of conceivable real rate disturbances in the U.S. economy has risen and so also has my estimate of the interest elasticity of money demand.

Suppose the U.S. economy in 1981 had been operating at a stable and low inflation rate and reasonably close to full employment. Suppose a disturbance had raised the real rate of interest by roughly the magnitude actually observed, and that as a consequence nominal interest rates had risen from 4 percent to 8 percent. This doubling of nominal rates is a logarithmic increase of 0.7, and multiplying .7 by an interest elasticity of money demand of 0.6 yields a logarithmic change in velocity of 0.42, or 50 percent. That is very large velocity disturbance, and one that would obviously have a considerable impact on both the real economy and the price level in the absence of a change in the money stock.

These calculations may provide a misleading picture of the magnitude of the problem because there is evidence that the real rate and the expected rate of inflation are negatively correlated (Fama and Gibbons, 1982). More casually, real rates fell in the late 1970s as inflation expectations and nominal rates rose, and real rates rose in the early 1980s as inflation expectations and nominal rates fell. In these two instances, then, inflationary expectations changed by more than the real rate so that the real rate and the nominal rate moved in opposite directions rather than in the same direction as assumed in the preceding paragraph.

The key issue here is whether the observed negative correlation between the real rate and inflation expectations is a accident of history or is embedded in the structure of the economy. The correlation almost certainly depends on whether real disturbances over a particular period originate primarily on the demand or the supply side of the economy and on whether a country has a fixed or floating exchange rate.[15]

To summarize, there is unmistakable evidence that a large real shock occurred in the early 1980s. There may also have been significant real shocks in the 1970s, although monetary shocks may be the larger part of the explanation. In any event, recent experience of inflation and disinflation in the United States suggests that monetary policy should be formulated in such a way as to deal with the possibility of large real disturbances. Put another way, there is more to monetary policy than simply avoiding monetary disturbances, as important as that aspect of policy must obviously be.

---

[15] For evidence on the behavior of the real exchange rate under fixed and floating rate regimes see Mussa (1986) and papers cited there.

Market Expectations Concerning Monetary Policy

Both interest rates and money growth became more volatile after the Federal Reserve changed its operating procedures in October 1979. After October 1979 market interest rates reacted quickly to weekly reports of money stock data every Thursday afternoon (Friday afternoon for part of the period under review). One useful outcome of the experience was a greatly improved understanding of market responses to monetary policy.[16]

To understand the issues here, it is best to think of Fed policy day by day as involving an adjustable peg for the federal funds rate. That description of Fed policy is certainly correct before October 1979, is approximately correct after October 1982, and is in dispute for the period between. If this model is accepted for the entire period, then market responses to receipt of information provide evidence on the market's evaluation of the factors that lead the Federal Reserve to adjust its federal funds rate peg.

Effects of Unexpected Money

Investigators have partitioned each week's money stock data release into expected and unexpected parts. The expected part is based on a survey of market participants conducted by Money Market Services every week prior to release of money stock data. The interest rate response has been measured over 24 hours spanning the announcement time, and over the period from 3:30 p.m. to 5:00 p.m. spanning the announcement time.

Although the data contain some anomalies, the basic finding is that the market responds to the unexpected part of the money stock data release. Roley and Troll (1983) and Roley (1986) estimated that an unexpected increase of $1 billion in the money stock raised the Treasury bill rate by a trivial 1.6 basis points from October 1977 to October 1979, by 10.4 basis points from October 1979 to October 1982, by 3.4 basis points from October 1982 to February 1984, and by 1.4 basis points from February 1984 to September 1985. The decline in the response is no doubt due to the fact that the Federal Reserve increasingly deemphasized money stock control after October 1982. My casual impression is that the response fell essentially to zero in 1986.

These estimates of the effects on the Treasury bill rate of an unexpected change in the money stock can also be used as an estimate of the effect of the expected change in the money stock. As information accumulates day by day the market will adjust its expectation of the change in the money stock (relative to the Federal Reserve's target change), and will presumably bid the bill rate up or down to the same extent as the documented effect from unexpected changes in money.

This discussion has assumed that the market's response to money stock data arises from speculation on future monetary policy rather than from the effects of money growth on the market's expectation of future inflation. This proposition is supported by the way in which the interest rate response to money growth changed over time as the Federal Reserve changed its operating procedures. However, the strongest evidence supporting this proposition arises from an examination of the response of the exchange rate to unexpected money. Engel and Frenkel (1984) studied data for October 1979 to August 1981 and showed that the dollar appreciated (depreciated) against the German mark in response to an unexpected increase (decrease) in the stock of money in the United States. This finding is consistent with the hypothesis that the interest rate response to unexpected money is a real rate response. If the inflationary anticipations argument were correct, the exchange rate should have depreciated in response to unexpected increases in the money stock.

---

[16] Sheehan (1985) provides a survey of findings and a bibliography. See also the paper by Deaves, Melino, and Pesando (1987).

I know of no study that sorts out the effects on interest rates of *actual* changes in the money stock as distinct from *announcements* of past changes. If I am correct that actual changes have little or no immediate effect on interest rates (the buffer stock argument), then rates were much more volatile after October 1979 both because the market's response coefficient to unexpected money was higher and because money growth was less stable and less predictable than it had been before. Thus, the market's increased response to announcements of money stock changes after October 1979 would not have created increased interest rate volatility had the Federal Reserve been successful in stabilizing money growth. The issue is not a minor one; the effect of the Fed's October 1979 procedures on interest rate volatility, which shows up dramatically in interest rate charts, had much to do with the Fed's abandonment of efforts to control money growth (Roley, 1982, p. 9; 1983 p. 10).

Although there is a dispute concerning the Fed's technical capacity to stabilize short-run money growth, there is no doubt whatsoever that the Fed can control nonborrowed bank reserves extremely precisely. The evidence strongly suggests that if the Fed were to pursue an announced policy of achieving a bank reserves target, and were to hit the target with a small margin of error, then the unexpected part of monetary policy would have minimal variance and interest rates would not be bid up or down by speculation on Fed actions to offset the error in hitting its announced target.

Effects of Other Information

There is also evidence of a dramatic change in October 1979 in the market's response to adjustments in the Federal Reserve's discount rate. Using data for the two-year period before and the three-year period after October 1979, Roley and Troll (1983) report that per 100 basis points increase (decrease) in the discount rate the bill rate rose (fell) 10 basis points in the first period and 54 basis points in the second period.

The best way to make sense of the observed effect of the discount rate on market rates is to assume that the Federal Reserve has persistent interest rate targets. Before October 1979 the Fed changed its federal funds rate target by a relatively small amount at a time, and a discount rate change announced to the market that the Fed had changed its funds rate target. After October 1979, the Fed was much more aggressive in changing its funds rate target and so the market's response to discount rate changes was much larger. On this interpretation, Fed policy after October 1979 did not involve a target path for bank reserves that was exogenous to short-run fluctuations in money market interest rates. Instead, Fed policy must have involved aggressive use of the federal funds rate as the policy instrument. The Fed adjusted the funds rate on the basis of the observed money stock and other data, and discount rate adjustments signalled changes in the target zone for the funds rate.

Roley and Troll (1983) and Roley (1986) also provide evidence on the effect on the Treasury bill rate of the unexpected component in official releases of data on inflation, unemployment, and industrial production. The estimated effects are all statistically insignificant except for the industrial production index in the period from February 1984 to September 1985. These results may understate the importance of data releases because monetary policy tends to be asymmetric. My hypothesis is that when inflation is the major concern of the public the Fed will tend to push up interest rates in response to reports of high inflation, high employment growth, and growing industrial production but will tend not to push interest rates down in response to reports in the opposite direction. Similarly, when recession is the major public concern the Fed will tend to push interest rates down in response to low inflation, falling employment, and falling production but will tend not to push rates up in response to reports in the opposite direction.[17]

Effects of Market Expectations on Policy

From this evidence it is clear that the market is extremely sensitive to the manner in which the Federal Reserve conducts policy. It should be no surprise that in the winter and spring of 1987 the market began to look especially closely at commodity price indexes as concerns developed that inflation might rise significantly. Recent appointees to the Board of Governors have suggested that commodity prices be used as a key indicator of the appropriate stance of monetary policy. On this view, monetary policy should become more restrictive when commodity prices are rising; what that means in practice is that the Fed should push money market interest rates higher.

These considerations raise a serious issue concerning how the Federal Reserve should interpret observed interest rate changes. Suppose, for example, that the Treasury bill rate rises in response to a report of higher commodity price inflation. If the Fed holds its target for the federal funds rate unchanged, and does so systematically in such cases, the market will learn that the Fed does not in fact change its funds rate target in response to commodity price information. In time, the bill rate will cease responding to commodity prices. At the opposite extreme, the Fed may ratify the market by adjusting the funds rate by an amount that matches the market's adjustment of the bill rate. In this case the Fed would simply follow the market, which would be an obviously unsatisfactory and ill-defined monetary policy.

---

[17] Another interesting piece of evidence on the market's response to monetary policy is in a recent paper by Mankiw, Miron, and Weil (1987). These authors examined the effect of the opening of the Federal Reserve System in November 1914 on the time series characteristics of interest rates and the term structure of rates. They conclude that the market responded quickly and in the direction expected from a priori knowledge about how the Fed would operate.

Suppose the Fed controls short-term interest rates and develops a systematic interest rate policy response to receipt of new information. Then, when the market changes long-term interest rates upon receipt of new information or revised forecasts of future data releases, the size of the rate change will reflect nothing more than the market's judgment of the Fed's response function. The Fed cannot then use long-term interest rates to provide information on disturbances elsewhere in the economy.

It is important to understand the problem here: it is *in principle* impossible to extract useful information about the private economy from any variable the Fed itself controls. No one would dream of trying to extract information about the private economy from the behavior of the discount rate. Nor should anyone try to extract information about the private economy from the behavior of the federal funds rate when that rate is closely controlled by the Federal Reserve, as was true before October 1979 and after October 1983.

Of course, longer-term interest rates reflect market expectations concerning *both* future data and the Fed's responses to the data. The point remains, however, that the Fed cannot extract from interest rates information about the economy without knowing market expectations concerning Fed policy. In fact, the principle holds for all speculative prices. The prices of storable goods—commodities, land, securities, and so on—reflect in part expectations of future monetary policy because speculators know that Fed actions will influence aggregate real demand and the general price level.

In sum, interest rates respond to disturbances in the real economy, to disturbances in money demand, and to changes in market expectations of future monetary policy. Recent studies have made clear that market expectations concerning policy have economically significant effects on interest rates. What this means is that a monetary policy based on controlling short-term rates cannot look at long-term rates to provide reliable information about the strength of the real economy and/or of inflationary expectations.

The Role of Monetary Policy in Reducing Inflation

Monetary policy played a critical role in reducing inflation in the early 1980s, but quantifying the contribution would go beyond the scope of this paper. The issues involve the direct contribution of lower money growth, the contribution of changed inflationary expectations, and the contribution of real disturbances.

Role of Lower Money Growth

In the late 1970s and 1980–81, inflation rose to a higher rate than would have been sustained in the long run by the rate of money growth at that time. On an annual average basis, $M1$ growth peaked at a little over 8 percent in 1978. Assuming secular growth of real GNP of about 3 percent per year, sustained $M1$ growth of 8 percent per year would yield secular inflation of 5 percent per year if velocity has zero secular increase and 8 percent inflation if velocity has a 3 percent secular increase. Actual inflation on an annual average basis as measured by the fixed weight GNP deflator was 9.3 percent in both 1980 and 1981.

Thus, 1980–81 inflation exceeded the steady rate inflation implied by 8 percent money growth by at least one and perhaps four percentage points. This excess inflation was presumably due to the *acceleration* of money growth in the late 1970s and/or the oil price shock of 1979. Ending the acceleration of money growth was part of monetary policy, but the effects of changing money growth should be distinguished from those of the rate of money growth itself. A transition to a sustained higher rate of money growth will, in the long run, raise nominal interest rates and velocity to higher equilibrium levels; while velocity is rising during the transition, inflation will outpace money growth. Conversely, velocity will fall during the transition to sustained lower rates of money growth and inflation.

Money growth fell in the early 1980s, and here again it is necessary to distinguish the effects of the deceleration from those of the lower rate itself. On an annual average basis, $M1$ growth fell to 6.5 percent in 1981, rose to 9.8 percent in 1983, and fell back to 6.0 percent in 1984. Inflation as measured by the fixed weight deflator fell to slightly less than 3 percent in 1986. Assuming secular real GNP growth of 3 percent per year and zero secular change in velocity, sustained $M1$ growth of 6 percent per year would yield sustained inflation of 3 percent per year. However, when the data are averaged over several years it is clear that in the 1980s inflation fell much more than did $M1$ growth.

On a qualitative level, the decline of inflation after 1980 is perfectly consistent with the deceleration of money growth but the magnitudes seem wrong. Although $M1$ growth did fall in 1980–81, its average rate over the 1979–84 period was essentially identical to its average rate over the 1975–79 period. It seems to me unreasonable to argue that monetary conditions alone can account for the extent and duration of the decline of inflation.

Role of Inflationary Expectations

One way to reconcile the sustained reduction of inflation from 1982–86 is to accept the hypothesis that inflation expectations declined substantially from their level in the late 1970s. Without question, inflation expectations did decline and the Federal Reserve's willingness to maintain a restrictive policy long into the 1981–82 cyclical contraction must have had something to do with the decline.

But it is difficult to attribute the durability of the decline in inflation expectations entirely to monetary policy.[18] It is illogical to argue that the market was simply relying on Fed chairman Paul Volcker, for it was unclear whether he would be reappointed in 1987 to another term as chairman, and what would happen in the event he was not reappointed. What was clear was that the Reagan administration had ceased to be concerned with inflation, as evidenced most clearly by the administration's policy starting in September 1985 to depreciate the dollar on the foreign exchanges. It was also clear that Congress and the administration both had an incentive to inflate to relieve pressures arising from the large federal budget deficit.

---

[18] From market commentary and the behavior of nominal interest rates, it appears that inflation expectations did not begin to rise again until early 1987.

In sum, it appears that significant weight must be accorded to nonmonetary conditions as shaping both the actual inflation rate and inflation expectations from 1981 through 1986. The obvious candidates for these nonmonetary conditions are those responsible for the high real rate of interest 1982–85.

Role of Nonmonetary Conditions

As mentioned earlier, there is substantial disagreement as to the causes of high real rates of interest in the United States 1981–85. However, there is agreement that the high real rates were primarily responsible for the real appreciation of the dollar, which reached its peak in February 1985. The strong dollar almost certainly had much to do with the rapidity of the decline of the overall inflation rate in 1982 and its further decline in 1986. Traded goods prices enter directly into price indexes, and also affect the prices of competing non-traded goods. Over the five years from 1981 to 1986 the fixed weight imports deflator fell by 9 percent while the fixed rate exports deflator rose by only 6 percent.

It is often argued that the decline of inflation in 1986 was due to an exogenous oil price decline, but this argument seriously underestimates the importance of the strong dollar. Consider in the abstract a commodity traded in world markets, and ask what happens when currency *A* appreciates in real terms against currency *B*. In general, the commodity's price should fall in terms of currency *A* and rise in terms of currency *B*, with the price changes depending on relative supplies and demands and their elasticities in countries *A* and *B*.

As the dollar appreciated after 1980, the OPEC oil cartel was able to maintain a relatively unchanged dollar price of oil, with the result that oil prices rose very substantially in non-dollar currencies. That outcome would not have occurred in a highly competitive oil market; however, neither was the outcome profit-maximizing for a cartel. Whether or not the OPEC cartel weakened in the 1980s, a decline in the dollar price of oil was a predictable consequence of substantial dollar appreciation. Given appreciation of the dollar, OPEC changed the timing of the decline of dollar oil prices but did not determine the decline itself.

In sum, there is good reason to believe that nonmonetary conditions contributed significantly to the reduction of U.S. inflation in the 1980s. However, it should be emphasized that these conditions were inherently transitory. The United States could not expect continuing *increases* in its real rate of interest and in the real value of its currency. Once these conditions stabilized, their contributions to reducing inflation would end and monetary fundamentals would dominate the outcome. Indeed, with lower real interest rates and sustained dollar depreciation starting in 1985, it is reasonable to expect that this process will work in the other direction in due time.

Implications of Recent Experience for Monetary Rules

After the Federal Reserve abandoned efforts to achieve money growth targets it returned by default to a policy of controlling short-term interest rates. There is no articulated theory of how such a policy should be managed. In practice, the Fed simply leans a little one way or the other depending on its view as to whether the economy needs more or less stimulus. Perhaps monetary policy is like riding a unicycle; a skilled practioner can do it without being able to explain just how. But everyone agrees that a pegged interest rate fails to provide a nominal anchor for the price level, and I continue to believe that it would be highly desirable if we could specify a monetary policy with more analytical content than the "tone and feel" of the expert practitioner. It still seems to me that the best opportunity for defining a good monetary policy lies in a rule built around control of money growth.

In past debates on money stock rules no clear distinction has been made between the application of a rule in a situation of stable initial conditions and the application in a period of high inherited inflation. Opponents of a monetary autopilot seem to assume that the economy is always like a ship in a crowded channel while proponents seem to assume that the economy is always on the high seas.

A Monetary Rule in the Steady State

Suppose money growth were constant at, say, 3 percent per year. What would be the operating characteristics of the economy once the transition to the rule was complete?

As argued earlier, recent experience suggests that the interest elasticity of the demand for money is high enough that real disturbances have a substantial potential to raise or lower the inflation rate and to push real growth temporarily above or below its equilibrium path. The relatively high elasticity also means, however, that independent disturbances to money demand will have less effect on nominal GNP in the short run than would be the case with a lower elasticity.

The buffer stock model of money demand requires a reinterpretation of past evidence on money demand disturbances. If it is true that these disturbances, as conventionally measured, are often dependent on disturbances elsewhere in the economy, then money demand behavior is far more stable and stabilizing, though less rigid, than many have come to believe. Conventionally measured money demand disturbances do not provide evidence supporting discretionary monetary policy designed to offset them; such disturbances often reflect money *supply* disturbances, which have effects of the type emphasized by rational expectations models.

Evidence on the behavior of interest rates under varying monetary policy operating procedures confirms the validity of the general point emphasized in the Lucas critique that parameters in macro model equations will not be stable because these equations are not truly structural. Estimated macro models do not provide any evidence whatsoever on how an economy would operate under a constant money growth rule, and that point will remain even if the money demand function in macro models is altered as I have suggested. My guess is that the only evidence that will be very persuasive will be gathered from actual experiments with a money growth rule because economists will not in the foreseeable future be able to extract truly structural equations from data generated under present monetary arrangements.

A Monetary Rule for Disinflation

The monetary rule most often advocated for bringing inflation down from an excessive rate is that of gradual, predetermined reductions in money growth. In the late 1970s a number of economists, myself included, advocated a gradualist monetary policy of this type, but I now believe that experience 1980–85 has demonstrated quite decisively that the gradualist prescription is unreliable. The decline of velocity was far greater than economists had predicted, and it seems very unlikely that the economy could have adjusted satisfactorily if money growth had been reduced by, say, one percentage point per year starting in 1980.

I have provided a more complete argument for this position earlier (Poole, 1986), but the reasoning behind the conclusion may be stated briefly. Start with a provisional assumption that under a policy of gradualism velocity in the 1980s would have followed the course actually observed. If money growth had declined gradually, then by 1985 nominal GNP would have been falling. Thus, if real growth had been positive in 1985, the price level would have been falling. Nominal interest rates would presumably have been lower than those actually observed, and so velocity would, if anything, have declined even more than its actual decline. It seems very unlikely that real growth could have been positive in these circumstances; the magnitude and speed of the adjustment from the inflationary era of the 1970s could not have proceeded rapidly enough to permit real growth.

This argument is very casual, and it depends almost entirely on my feel for the data, which I believe many others share in this case. A less casual observation, though, is that there is certainly no precedent suggesting that the real economy could have adjusted without a recession much deeper and more prolonged than the recession actually experienced. Under the scenario of the preceding paragraph the decline of the inflation rate between 1980 and 1985 would have been greater than the decline of inflation between 1928 and 1932.[19]

There is a serious and probably insurmountable problem to designing a predetermined money growth path to reduce inflation. When policymakers make a commitment to lower inflation they should expect that velocity growth will decline, and perhaps become negative, if the policy is successful. The pace of adjustment in financial markets and the real economy will depend on expectations concerning future monetary policy and future inflation. If the market expects lower inflation, then inflation will decline without a decline in money growth. Under these circumstances, a decline in money growth might have a significant depressing effect on real activity. A recession is undesirable in and of itself, but also may make it impossible politically to sustain the policy.

---

[19] Using annual average data for the Consumer Price Index, the inflation rate was −1.3 percent in 1982 and −10.3 percent in 1932, or a deceleration of 9.0 percentage points. In 1980 the inflation rate was 13.5 percent and in 1985 the rate was 3.6 percent, or a deceleration of 9.9 percentage points. Under the gradualism scenario the deceleration of inflation would have been greater. Different price indexes would yield somewhat different numbers but would not alter the conclusion.

The dilemma is that the necessary reduction of inflation expectations probably requires that money growth fall. It is irrational for anyone to bet on lower inflation on the basis of a central bank's promises with no evidence that the central bank is reducing money growth. Indeed, a recession may be necessary to provide the evidence that the central bank is serious. Are the regularities of behavior of leaders of democratic societies such that they are not believed on important issues unless they expend political capital in pursuit of their policies? A constitutional amendment might avoid this problem, but at a time of great public concern about inflation could an economist make a successful case for an amendment that merely required the central bank to stabilize money growth for, say, four years before starting to reduce it?

Consider the following numbers. Using annual data and continuous compounding, $M1$ growth was 7.5 percent per year 1976–79 and 7.6 percent per year 1979–83. In this second period money growth was first lower, which assisted the Federal Reserve in establishing credibility, and then higher. From 1983 to 1986 $M1$ growth was higher yet—an average annual rate of almost 11 percent. How much of this burst of money growth is part of a continuing velocity adjustment to lower inflation and lower interest rates and how much is the leading wave of the next acceleration of inflation?

To achieve the benefits of a constant money growth rule in the steady state there must at some point be a decision that the transition period of managed money during disinflation should be brought to an end. Just as I have come to believe that there is no satisfactory rule for managing monetary policy during the transition, I also believe that there is no formula to determine when the transition should terminate. Paul Volcker has earned a place in history as an extraordinarily skilled manager of the disinflation process, but his influence is unlikely to be permanent because he has not established a monetary rule or standard of some sort to guide his successors.

In summary, the last ten years has been a period rich in monetary policy lessons. The money demand function has displayed characteristics different from those that seemed so well established. Real interest rate in the 1980s were at sustained levels previously associated with depression, and yet the economy grew at a rapid rate after the 1981–82 recession. Recent experience makes clear that market interest rates are determined very importantly by expectations concerning future monetary policy. Although events have not changed the case for a monetary rule of constant money growth in the steady state, they have not been kind to the rule of gradual reductions of money growth to slow inflation.

■ *Special thanks go to Kerry Reinertsen and Louis Chan for research assistance, to Data Resources, Inc. for providing data, and to editors of this journal for their careful comments on an earlier draft of the paper. Of course, responsibility for errors remains with the author.*

References

Bordo, Michael, and Lars Jonung, "The Global Velocity Curve 1952–1982," National Bureau of Economic Research Working Paper No. 2074, November 1986.

Campbell, John Y., and Richard H. Clarida, "The Dollar and Real Interest Rates," National Bureau of Economic Research Working Paper No. 2151. Forthcoming in Brunner, Karl, and Allan H. Meltzer, eds., Carnegie-Rochester Conference Series on Public Policy, Amsterdam: North Holland.

Cooley, Thomas, F., and Stephen F. LeRoy, "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

Deaves, Richard, Angelo Melino, and James E. Pesando, "The Response of Interest Rates to the Federal Reserve's Weekly Money Announcements: The 'Puzzle' of Anticipated Money," National Bureau of Economic Research Working Paper No. 2125, January 1987.

Engel, Charles E. and Jeffrey F. Frankel, *Journal of Monetary Economics*, January 1984, *13*, 31–39.

Fama, Eugene, F., "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

Fama, Eugene, F., and Michael R. Gibbons, "Inflation, Real Returns and Capital Investment." *Journal of Monetary Economics*, May 1982, *9*, 297–323.

Fisher, Irving, *The Purchasing Power of Money*. New York: Augustus M. Kelley, 1963 (reprint of second revised edition 1922).

Friedman, Milton, and Anna J. Schwartz, *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University for the National Bureau of Economic Research, 1963.

Goldfeld, Stephen M., "The Case of the Missing Money." In Okun, Arthur M., and George L. Perry, eds., *Brookings Papers on Economic Activity*, 1976:3, pp. 683–730.

Goldfeld, Stephen M., "The Demand for Money Revisited." In Okun, Arthur M., and George L. Perry, eds., *Brookings Papers on Economic Activity*, 1973:3, pp. 577–638.

Goodfriend, Marvin, "Reinterpreting Money Demand Regressions," In Brunner, Karl, and Allan H. Meltzer, eds., *Understanding Monetary Regimes*. Carnegie-Rochester Conference Series on Public Policy, 22, Amsterdam: North Holland, 1985, pp. 207–42.

Gould, John P., and Charles R. Nelson, "The Stochastic Structure of the Velocity of Money," *American Economic Review*, June 1974, *64*, 405–18.

Huizinga, John, and Frederic S. Mishkin, "Monetary Policy Regime Shifts and the Unusual Behavior of Real Interest Rates." In Brunner, Karl, and Allan H. Meltzer, eds., *The National Bureau Method, International Capital Mobility and Other Essays*. Carnegie-Rochester Conference Series on Public Policy, 24, Amsterdam: North Holland, 1986, pp. 231–74.

Judd, John P., and John L. Scadding, "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature*, September 1982, *20*, 993–1023.

Laidler, David, "The Demand for Money in the United States—Yet Again." In Brunner, Karl, and Allan H. Meltzer, eds., *On the State of Macroeconomics*. Carnegie-Rochester Conference Series on Public Policy, 12, Amsterdam: North Holland, 1980, pp. 219–71.

Laidler, David, "Expectations and Adjustment in the Monetary Sector Revisited: A Comment." In Brunner, Karl, and Allan H. Meltzer, eds., *Understanding Monetary Regimes*. Carnegie-Rochester Conference Series on Public Policy, 22, Amsterdam: North Holland, 1985, pp. 234–54.

Lucas, Robert E. Jr., "Econometric Policy Evaluation: A Critique." In Brunner, Karl, and Allan H. Meltzer, eds., *The Phillips Curve and Labor Markets*. Carnegie-Rochester Conference Series on Public Policy, 1, Amsterdam: North Holland, 1976, pp. 19–46.

Mankiw, N. Gregory, Jeffrey A. Miron, and David N. Weil, "The Adjustment of Expectations to a Change in Regime: A Study of the Founding of the Federal Reserve," National Bureau of Economic Research Working Paper No. 2124, January 1987.

Miller, Merton, H., and Daniel Orr, "A Model of the Demand for Money by Firms," *Quarterly Journal of Economics*, August 1966, *80*, 413–35.

Mussa, Michael, "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications." In Brunner, Karl, and Allan H. Meltzer, eds., *Real Business Cycles, Real Exchange Rates, and Actual Policies*. Carnegie-Rochester Conference Series on Public Policy, 25, Amsterdam: North Holland, 1986, pp. 117–213.

Nelson, Charles R., and Charles I. Plosser, "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Poole, William, "Federal Reserve Operating Procedures: A Survey and Evaluation of the Historical Record since October 1979," *Journal of Money, Credit and Banking*, November 1982, *14*, 575–96.

Poole, William, "Is Monetarism Dead?," *Business Economics*, October 1986, 10–15.

Poole, William, "Whither Money Demand?" In Okun, Arthur M., and George L. Perry, eds., *Brookings Papers on Economic Activity*, 1970:3, pp. 485–500.

Rasche, Robert H., "*M1* Velocity and Money Demand Functions: Do Stable Relationships Exist?" In Brunner, Karl, and Allan H. Meltzer, eds., Carnegie-Rochester Conference Series on Public Policy, 27, Amsterdam: North Holland, 1987, pp 9–88.

Roley, V. Vance, "Market Perceptions of U.S. Monetary Policy Since 1982," Federal Reserve Bank of Kansas City *Economic Review*, May 1986, 27–40.

Roley, V. Vance, "Weekly Money Supply Announcements and the Volatility of Short-Term Interest Rates," Federal Reserve Bank of Kansas City *Economic Review*, April 1982, 3–15.

Roley, V. Vance, and Rick Troll, "The Impact of New Economic Information on the Volatility of Short-Term Interest Rates," Federal Reserve Bank of Kansas City *Economic Review*, February 1983, 3–15.

Schwert, G. William, "The Time Series Behavior of Real Interest Rates." In Brunner, Karl, and Allan H. Meltzer, eds., *The National Bureau Method, International Capital Mobility and Other Essays*. Carnegie-Rochester Conference Series on Public Policy, 24, Amsterdam: North Holland, 1986, pp. 275–88.

Sheehan, Richard G., "Weekly Money Announcements: New Information and Its Effects," Federal Reserve Bank of St. Louis *Review*, August/September 1985.

Wilcox, James A., "Why Real Interest Rates Were So Low in the 1970s," *American Economic Review*, March 1983, *73*, 44–53.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 101–115

Illegal Immigration and Immigration Control

Barry R. Chiswick

The Immigration Reform and Control Act (IRCA) was passed in the closing days of the 1986 legislative session. The primary purpose of IRCA is to remove illegal aliens from the U.S. labor market. It has two primary policy instruments. One is granting legal status or amnesty for certain illegal aliens, thereby in part "wiping the slate clean." The other is imposing penalties, referred to as employer sanctions, against employers who "knowingly" hire illegal aliens. Employer sanctions are intended to reduce the demand for illegal alien labor.

IRCA was the culmination of a decade of activity by the White House, Congress, a Select Commission and numerous task forces. The basic outline of IRCA—amnesty and employer sanctions—had been endorsed by the Ford, Carter and Reagan administrations and had had considerable but not sufficient support in Congress for a decade. Some interest groups supported amnesty and opposed employer sanctions. These included Hispanic organizations, many growers and some civil liberties groups. Others opposed amnesty and supported employer sanctions because of their concerns with the impact of the growing population of the country on the environment or the economic impact of low-skilled migrants. Still others opposed both amnesty and employer sanctions because the legislation was deemed inadequate to address fundamental issues. Yet others, including the primary sponsors Senator Simpson and Congressmen Mazzoli and Rodino, favored both amnesty and sanctions so as to fully incorporate those illegal aliens already established in the United States, while closing the door to further illegal immigration by eliminating the jobs magnet. The legislative stalemate was finally broken only when amendments were incorporated to assure the

■ *Barry R. Chiswick is Research Professor, Department of Economics and Survey Research Laboratory, University of Illinois at Chicago. This article was written while he was John M. Olin Visiting Professor, Center for the Study of the Economy and the State, University of Chicago.*

agricultural interests that the new law would not reduce the supply of farm labor, especially during the harvest seasons.

The first sections of this paper develop an economic analysis of the illegal alien labor market, including the determinants of illegal migration and the impact on the economy. Then, that model is applied to the major provisions of IRCA and used to describe its likely consequences. A concluding section argues that because IRCA does not address the economic realities it is not likely to accomplish its objectives. The partial amnesty and impotent employer sanctions have not solved the illegal alien dilemma.

What Is Illegal Immigration?

By definition, illegal immigration arises from a divergence between whom the United States will accept as an immigrant and the desire of some foreign nationals to live and work in this country. Illegal immigration is as old as U.S. immigration law. The first immigration restrictions were imposed on the Chinese in the late 19th century, and the first illegal aliens were therefore Chinese. With changes in U.S. immigration law and economic and political conditions in the United States and other countries, the nature and characteristics of illegal immigration have also changed.

In recent years the United States has accepted nearly 600,000 immigrants and refugees annually. This is approximately one immigrant per year for every 400 residents of the country. By way of contrast in the peak immigration period (1905 to 1914) the U.S. accepted over one million immigrants annually, or one for every 83 residents in 1905.

Under current law, nearly all legal immigrants are either refugees or the relatives of a U.S. citizen or resident alien (that is, an immigrant who has not become a naturalized citizen). By "relative" is meant the spouse, parents and children of citizens and resident aliens, and siblings of U.S. citizens. There are numerical limits on certain categories of immigrants sponsored by U.S. relatives. A small number of individuals, about 22,000 per year, receive an immigrant visa on the basis of their own occupation or skills. The spouse and minor children of one who qualifies for a visa generally also receive visas. Many individuals in other countries wish to immigrate but either cannot do so legally or do not wish to wait in the queue. Some of these become illegal aliens.

Illegal aliens are foreign-born persons whose mere presence in the United States is in violation of the law or who have violated a condition of a lawful entry. They are primarily persons who entered the country surreptitiously by evading an immigration service inspection. Those who "enter without inspection" (referred to as EWIs) are by far the most numerous. They are disproportionally Mexican nationals who walk or ride across the southwestern border. A smaller group of illegal aliens are visa abusers. They are primarily from eastern hemisphere countries and either work in violation of a student or tourist visa or have remained in the United States longer than is permitted by the visa. It is generally easier for eastern hemisphere nationals who aspire to work in the United States illegally to enter with a student or tourist visa, even if the valid visa is obtained fraudulently, than to attempt surreptitious entry.

Table 1

Legal and Illegal Immigration to the United States by Country of Origin

(in percent)

			Illegal Aliens		
				Apprehensions	
Origin	*Foreign Born Population*[b] *(1980)*	*Legal Immigration*[c] *(1985)*	*Resident Population*[d] *(1980)*	*Total*[c] *(1985)*	*District Offices*[c] *(1985)*
Europe	36.6	11.1	7.3	0.5	6.3
Asia	18.0	46.4	10.4	0.7	8.7
North America[a]	33.1	31.9	71.9	97.4	71.8
Mexico	15.6	10.7	55.0	93.9	57.5
Canada	6.0	2.0	1.2	0.5	1.3
South America	4.0	6.9	6.2	0.8	6.7
Other, and Country Not Reported	8.3	3.7	4.1	0.5	6.5
Total	100.0	100.0	100.0	100.0	100.0
Number in Thousands	14,079.9	570.0	2,057.0	1,348.7	83.9

[a] Includes Central America and the Caribbean.

[b] Based on 1980 Census of Population. The data include about 2 million illegal aliens (see column 3).

[c] Immigration and Naturalization Service administrative records. Legal immigrants refers to those given permanent resident alien status in fiscal year 1985. District Office apprehensions refer to interior enforcement.

[d] Estimate of the number of illegal aliens responding to the 1980 Census of Population. Precision of estimates is lower for the country of origin categories than for the total. The number of Census respondents is about half of the total number of illegal aliens estimated elsewhere (Siegel, *et al.*, 1981).

Sources: U.S. Bureau of the Census, *Statistical Abstract of the United States, 1987*, 107th edition, Washington, D.C.: U.S. Government Printing Office, 1987, p. 11 (Table 7), p. 37 (Table 42).

Jeffrey S. Passel and Karen A. Woodrow, "Geographic Distribution of Undocumented Immigrants: Estimates of Undocumented Aliens Counted in the 1980 Census by State." *International Migration Review*, 18, Fall 1984, 642–71.

U.S. Department of Justice, *1985 Statistical Yearbook of the Immigration and Naturalization Service*. Washington, D.C.: U.S. Government Printing Office, 1986, p. 177 (Table ENF1.2) and pp. 178–79 (Table ENF1.3).

Among those unable to migrate legally, costs and benefits are important in determining who becomes an illegal alien. Migration costs may play a more important role for illegal than for legal migration. Several attempts may be needed to penetrate the border successfully and the possibility, however slim, always exists of being apprehended and deported once one has become established in the United States. Furthermore, the risks of detection often rise with the number of family members who join the illegal alien, which encourages leaving dependents in the country of origin. This family separation stimulates more frequent trips back home and increases the temporary nature of much illegal migration.

These considerations help to determine which countries account for the greatest numbers of illegal aliens to the United States. Although about 10 percent of the legal immigrants are Mexican nationals, as shown in Table 1, it is estimated more than half of the resident illegal alien population is of Mexican origin. About 90 percent of apprehended illegal aliens are Mexican nationals. However, the data on apprehensions distort the profile of illegal aliens because of the "revolving door" along the U.S.-Mexican border. It is relatively inexpensive to attempt to cross the border, and if apprehended during the attempt, to try again the next day. Moreover, it is far more difficult for the immigration authorities to apprehend an illegal alien already in the United States, whether through a surreptitious entry or by violating a lawfully obtained visa.[1] The careful scrutiny given visa applicants, which offends many foreign students and visitors to the United States, is intended to ferret out those most likely to violate their visas.

Skill levels also provide differential incentives for illegal migration (Chiswick, 1988). Individuals with occupational characteristics that would reveal their illegal status, such as the need to acquire a license or show other credentials, are less likely to migrate. The large wage gap between the United States and Mexico is larger for those with lower levels of skill and this may generate a higher rate of return from migration to the U.S. for the low-skilled workers. In addition, family separation which increases to-and-fro migration discourages the illegal migration of workers with firm-specific and country-specific skills, and discourages the acquisition of these skills by those who become illegal aliens. Evidence from Mexico indicates that it is not the least skilled and the poorest in Mexico who migrate to the United States; some skill and some resources are needed for a successful northward migration (Cross and Sandos, 1981). However, those who do migrate are very low skilled by U.S. standards.

Thus, the combined effects of the desire to avoid detection and the greater temporary nature of the migration has an effect on the skill level of those who become illegal aliens. Since the dollar magnitude of country-specific skills tends to rise with the overall skill level of workers, even if it declines as a share of total skills, illegal immigration is relatively more attractive for low-skilled workers. Data indicate that illegal aliens do in fact have lower levels of formal schooling, fewer skills acquired on the job, and a lower occupational status than have legal immigrants from the same country (for example, Massey, 1987). The gap between the skills of legal and illegal aliens is smaller for countries whose legal immigrants are low skilled, and the gap rises with the skill level of the legal immigrants (Chiswick, 1984).

It is widely believed that illegal immigration increased sharply in the 1970s and 1980s. Although no hard-and-fast numbers exist, Siegel et al. (1981) estimated that from a total illegal population of a few hundred thousand in the United States in the late 1960s, employed primarily on a seasonal basis in agriculture, the number of illegal aliens living in the United States grew to between three and six million by

---

[1] Of the 1.3 million apprehensions in fiscal year 1985, only 84,000 (6.2 percent) were by INS District Office inspectors (interior enforcement), the remainder were by the Border Patrol, primarily at or within a few miles of the border or a port of entry. Although Mexican nationals comprise 94 percent of the apprehensions, they were only 58 percent of District Office apprehension (see Table 1). At the time of apprehension, nearly 153,000 aliens (12 percent) reported they were employed, 52 percent in agriculture, and 48 percen in non-agricultural jobs. For additional data on apprehended aliens as well as other data from the IN administrative records, see the *Statistical Yearbook of the Immigration and Naturalization Service* published by th Department of Justice.

1980, living primarily in urban areas. The number of illegal aliens has increased since then.

The sharp increase in the number of illegal aliens in the past two decades had several causes. One was the end of the *bracero* program that had been in effect from 1942 to 1964 to supply low-skilled Mexican workers to harvest fruits and vegetables in the United States. Many legal workers became illegal aliens because the change in the law was not accompanied by a change in economic incentives for Mexican workers and the U.S. growers.

But more important to the overall increase of illegal immigration has been the increase in the supply of immigrants. Even though the annual number of legal immigrants doubled in the past 25 years, the supply of immigrants to the United States grew even faster. Some of the increased supply of immigrants is due to declines in the cost of transportation, communication and information. It is also due to changing economic and political circumstances in the countries of origin. Poor job opportunities for the growing youth population, combined with the resources to finance illegal migration, have served as a stimulus. For example, during the 1960s and early 1970s Mexico's high fertility rate resulted in a natural rate of population increase of about $3\frac{1}{2}$ percent per year. Although the fertility rate has declined sharply (from 43 births per 1,000 population in 1973 to 32 births per 1,000 population in 1986), and the natural rate of growth of the population has slowed to $2\frac{1}{2}$ percent per year, the large cohorts born in the 1960s are now the young adult members of the Mexican labor force. Mexico's total population is now 82 million.

Mexican nationals, in particular, have a very low cost of migration, can penetrate the border easily even if it requires more than one attempt, and have a very low probability of being detected and deported once they have penetrated the border. Thus, their cost of illegal migration is low.

The benefits of migration include the improved training opportunities and the much higher real wages obtained in the United States, even if the wages seem low by U.S. standards. Although it is difficult to make international wage comparisons, real U.S. GNP per capita is more than six times the Mexican level. Among a sample of illegal aliens apprehended in 1974, nearly 40 percent did not have running water and about a quarter did not have electricity in Mexico, compared to 5 percent and 2 percent, respectively, in the United States (Chiswick, Ch. 2, 1988). A study of Mexican illegal aliens working in the 1978 harvest season in Oregon estimated that they earned six times what they could have earned in Mexico, and even after deducting the costs of the seasonal migration (including payments to alien smugglers) and certain additional expenses for living in the United States, their net U.S. earnings were three times their Mexican alternative (Cuthbert and Stevens, 1981). As has been shown for the United States (Wright, 1987), large regional wage differentials sometimes persist far longer than might be implied by simple models of factor price equalization. The combination of high benefits and low costs provides a powerful stimulus for illegal migration from Mexico.

It is difficult to do research on the demographic and economic characteristics of a population that has an incentive to avoid being detected, not only by the enforcement authorities but also by survey researchers. As a result, knowledge of the illegal alien population is based largely on piecemeal studies of selected segments of the population. Enough such studies have been done, however, to suggest a profile for illegal aliens (see, for example, Chiswick, 1988, and the references therein).

The statistical evidence does not indicate that illegal aliens receive wages that are lower than those of other workers with comparable levels of skill (for example, Chiswick, 1988; Massey, 1987). Thus, the low wages of illegal aliens are generally the consequence of their skill levels and not of "exploitation." While in the United States, illegal aliens apparently undergo considerable job mobility as they adapt to the U.S. labor market, and this mobility limits the extent to which they can be exploited. As has been found for legal immigrants, their average earnings improve sharply with experience in the U.S. labor market, and especially according to their experience with their current U.S. employer. To some extent this reflects employment lasting longer the better the "match" (and the higher the wage). It also reflects, however, the favorable effect of acquiring U.S.-specific and firm-specific skills.

Economic Impact of Illegal Migration

Illegal immigrants have an impact on the economic well-being of the legal resident population of the United States. The impact is on both the level of income and the distribution of income.

For simplicity, consider a three-factor model of the economy.[2] There are two types of workers, low-skilled and high-skilled, and within each type all workers are homogenous. The third factor of production is physical capital, which is in fixed supply. In this model, assume that the three factors are substitutes for each other and as a first approximation the production function has a constant elasticity of substitution. The impact of illegal migration depends on the characteristics of the aliens. As has been discussed above, illegal aliens are disproportionately low-schooled, inexperienced, non-English-speaking workers. Even though their skills, occupational status, and earnings improve during a period of residence in the United States, most remain relatively low-skilled workers.

The entry into the labor force of additional low-skilled workers (whether due to illegal migration, legal migration or other causes) depresses the marginal productivity of all low-skilled workers. The decline in the marginal product of low-skilled natives generally translates into lower wages. However, it may also lead to a decrease in the employment of native workers, if low-skilled native workers have an upward-rising labor supply curve or if there is downward wage rigidity, say, because the market wage is bounded by the legal minimum wage or a union-imposed wage.

The increased immigration of low-skilled workers will lead to an increase in the marginal productivity and hence the wage rate of higher-skilled workers and the return on capital. This effect arises from the principle of complementarity: the productivity of a factor increases the greater the amount of other factors with which it works. Although one native factor loses and the other native factors gain, the overall income of the native population increases; the losses to native low-skilled labor are more than offset by the gains to native high-skilled labor and capital (Chiswick, 1982).

---

[2] For a fuller discussion of the model and its implications for the economic impact of immigrants, see Chiswick (1982).

The average income of the native population, as well as of the low-skilled immigrants, has increased. However, using conventional measures that do not account for changes in the composition of the population, the average income of the total population (natives and immigrants) in the immigrant-receiving country has declined. More importantly, perhaps, the inequality of income in the destination country has increased: there are more low-skilled workers, and the low-skilled workers have a lower average income.

In the 19th and early 20th centuries there were virtually no public transfer programs and little concern for the distribution of income. There was, therefore, little public policy interest in the distributional impacts of the massive low-skilled immigration that occurred at that time and the United States could easily maintain a largely open-door immigration policy. Today, as a result of considerable public concern over both income inequality and poverty, numerous large income transfer programs have been enacted. Yet by statute and court decisions, the United States generally allows legal immigrants the same access to income transfers and social services as natives.[3]

Because the aggregate income of the native population in the destination country has increased, it is in principle possible to achieve Pareto optimality if the immigrants do not receive benefits from the income transfer system. Through taxes and transfers some of the gains of high-skilled workers and the owners of capital can be transferred to low-skilled native workers to make all natives at least as well off as before the low-skilled migration. Thus, the low-skilled natives can, in principle, be "held harmless" (that is, kept at their previous level of income) in spite of the low-skilled immigration because the adverse impact on their wages and employment can be mitigated by paying lower taxes and receiving additional benefits from unemployment compensation, food stamps, AFDC, and other programs.[4] For this compensation to be possible, of course, the adverse effects of taxes and transfers on labor supply and capital formation must be sufficiently small.

Suppose, however, the income transfer system "holds harmless" low-skilled native workers and treats immigrants and natives equally. Then, in the framework of this model, the immigration of low-skilled workers would mean that the aggregate income transfers would exceed the increase in income of high-skilled native workers and capital. Under these circumstances, the inclusion of low-skilled aliens in the income transfer system means that a Pareto optimal transfer cannot take place. Thus, low-skilled immigration can make the native population as a whole worse off.

---

[3] Perhaps the most notable exception prior to IRCA is that since the 1980 Social Security Amendments Supplemental Security Income (which provides benefits for the aged and disabled who would otherwise be in poverty) cannot be received by immigrants within their first three years in the United States, unless the disability arises after immigration.

[4] Many of the current recipients of income-contingent transfers, particularly Aid to Families with Dependent Children (AFDC) and Supplemental Security Income (SSI), have little or no labor market attachment and hence do not suffer a direct adverse impact.

This poses a policy dilemma. The immigration of low-skilled workers with their dependents may reduce the income of the native population net of the tax-transfer systems; their contribution can be made positive if they come without dependent family members or they are denied access to income transfers and social services. Yet few in the United States would want within our borders a legally sanctioned permanent or long term "second-class" population that cannot bring immediate relatives with them or whose children in the United States grow up in poverty.

However, since the end of the *bracero* program in 1964, the U.S. approach to illegal migration has inadvertently had many features of an "optimal" policy, given certain institutional constraints. The United States received on a temporary basis large flows of low-skilled young adult workers for urban and rural jobs. These workers generally came without dependent family members, and they did not receive U.S. income transfers and social services. Indeed, it is often said that many paid payroll taxes for social insurance programs without having realistic access to benefits from these programs.

The seemingly optimal policy of "tolerating" illegal immigration turned out to be a short-lived solution. While most illegal aliens, particularly from Mexico, viewed themselves as temporary migrants, with the passage of time and more recent arrivals, an increasing number remained in the United States. As the European guest worker experience also suggests, it seems difficult for liberal democracies to prevent temporary migration from evolving into permanent migration. As these migrants stayed longer in the United States, their attachment to the U.S. labor market and society became stronger. With greater permanence came a decline in the probability of being deported—there is learning-by-doing in avoiding detection by the immigration authorities. As residence in the United States becomes more permanent and the likelihood of deportation declines, an increasing number bring dependents or develop families while living here. The result is the emergence of a primarily low-skilled, low-income population that either lives outside the legal system or has access to the income transfers. Neither of these outcomes is desirable.

Provisions of the Immigration Reform and Control Act

The key features of the IRCA, amnesty and enforcement attempt to address many of the problems created by the uncontrolled illegal immigration (U.S. House of Representatives, 1986).

Amnesty and Labor Supply

IRCA grants amnesty to certain categories of illegal aliens. The purpose of amnesty is to "wipe the slate clean" so that other policies can be implemented to reduce illegal immigration. IRCA also provides an assured flow of low-skilled workers for fruit and vegetable agriculture. Aliens granted legal status will not be subject to, nor will they be charged against, the various numerical limits in U.S. immigration law. They therefore represent a one-time increase in the legal alien population that in the long run is likely to increase legal immigration above the current level of about 600,000 a year because of the favorable treatment granted immediate family members in current immigration law.

Provisions of the Amnesty. Those who could demonstrate they have resided continuously in the United States in an illegal status since before January 1, 1982, and who filed their petitions during a 12-month period beginning May 5, 1987, were eligible for legalization. Illegal aliens who entered after January 1, 1982, are excluded on the grounds that they are not so firmly established that they cannot leave. The requirement that there be "continuous" residence in an illegal status is defined to permit periodic, perhaps even lengthy, trips outside the United States. Inexplicably, aliens who have alternated periods of legal and illegal status in the U.S. since January 1, 1982, could not qualify for amnesty. The provisions have been written to generally bar from amnesty foreign students who have been working in the United States in violation of their student visas.

The qualified aliens are granted "temporary resident alien" status which permits living and working within the United States and travel abroad. After 18 months in the temporary category, the aliens enter a 12 month period during which they *must* apply for permanent resident alien (immigrant) status. If they fail to apply and satisfy certain quite simple English language proficiency requirements, they revert to an illegal status. Like other immigrants, legalized aliens may apply for citizenship after five years as permanent resident aliens.

Under a separate Special Agricultural Worker (SAW) program in IRCA, illegal aliens who have worked in U.S. perishable crop agriculture for at least 90 days in the past year ending in May 1986 could qualify for legal status. The required duration in the temporary resident alien category is shorter for those who have worked in U.S. seasonal agriculture for up to three years.

IRCA also provides for another category of foreign agricultural workers, referred to as Replenishment Agricultural Workers (RAWs). Beginning in fiscal year 1990 and lasting for three years, replenishment workers may be admitted if there is a "shortage" of workers in perishable crop agriculture. A very broad definition of "perishability" is being written into the regulations. Illegal alien workers have not been used in harvesting the most perishable of crops because of the greater uncertainty of their labor supply. The number of replenishment workers is to be determined annually by the Department of Agriculture, but may not exceed the number of SAWs given legal status.

The replenishment worker program is a large-scale agricultural guest-worker program, but unlike the *bracero* program, the aliens may work in non-farm jobs when it is not the harvest season. Also unlike most temporary worker programs, replenishment workers are granted very favorable treatment for acquiring immigrant status. They must work on perishable crops for at least 90 days in each of three consecutive years to maintain their legal status and to qualify for permanent resident alien status.

The SAW and RAW programs were added to IRCA to enable the legislation to attract sufficient support among farmers to assure passage. To further protect the growers, enforcement of immigration law on farms and ranches was made more difficult. The law requires, for the first time, that the Immigration and Naturalization Service (INS) obtain a search warrant before it can conduct a raid on the open areas of a farm or ranch. These provisions have encouraged some to dub IRCA "the Growers Protection Act of 1986."

Applicants for Amnesty. The INS originally estimated that up to 4 million illegal aliens would be legalized under the regular program and several hundred thousand under the SAW program. These initial estimates were based on exaggerated estimates of the number of illegal aliens in the U.S. By the time the application period began, the estimate for the regular program was reduced to 2 million. The final data indicate that there were $1\frac{1}{2}$ million applicants in the regular program and one-half million under the SAW program.

Among the regular program applicants, 71 percent were from Mexico, 8 percent from El Salvador, nearly 3 percent from Guatemala and about one percent each from the Philippines, Columbia, Haiti, Nicaragua, Poland, Iran and the Dominican Republic. All other countries contributed the remaining 11 percent. In contrast, under the SAW program 81 percent were from Mexico, nearly 6 percent were from Haiti and all other countries accounted for the remaining 13 percent. Combining the two legalization programs, there were 200 or more applicants from 93 separate countries, representing every corner of the world. The preliminary approval rate has been 93 percent under the regular program and 87 percent under the SAW program.

In the regular program 55 percent of the applicants were male and 70 percent were 20 to 44 years of age (median age 30 years), in contrast to 83 percent male and 82 percent prime age (median age 28 years) in the agricultural program. The regular program applicants who reported an occupation were primarily laborers or service workers; the agricultural program participants were primarily engaged in fruit, vegetable and melon agriculture.

The primary states of residence of the applicants under the regular program were California (54 percent), Texas (19 percent), Illinois (7 percent), and New York (6 percent). For the agricultural program they were California (also 54 percent), Florida (14 percent), Texas (7 percent) and Arizona (5 percent).

Consequences of the Amnesty. Many eligible illegal aliens did not come forward because they have dependent family members living with them in the United States who cannot satisfy the residency requirement. Others have lived in the United States since before 1982, but cannot satisfy the requirement of "continuous" residence. Still others alternated periods of legal and illegal status and therefore have not been in a continuous illegal status. Other illegal aliens have already acquired *de facto* amnesty and have obtained sufficient documentation to live and work in the United States. For them the application fees may not be worth the benefits from *de jure* legal status. Some illegal aliens are fearful that they will be required to pay back taxes for unreported income earned in previous years. Finally, many illegal aliens mistrust the immigration authorities and are fearful of deportation.

The amnesty and replenishment worker provisions had two objectives: to protect illegal aliens established in the United States from the consequences of employer sanctions and to continue the supply of low-skilled workers for agriculture. The likely consequences, however, go beyond that. It is clear that many aliens will remain in an illegal status.

For several reasons, IRCA's provisions will increase the supply of low-skilled labor in the U.S. labor market, although this was not the intent of the legislation. First, the now-legalized aliens will more readily receive unemployment compensation in the off-season, discouraging temporary exits. Second, dependents will be able to migrate lawfully to the United States, further increasing the labor market attachment of the illegal alien workers. Third, many of the dependents will, if not immediately then some time later, join the labor market. These dependents of low-skilled legalized aliens are also likely to be low-skilled workers.

Fourth, amnesties can be effective for solving what are perceived to be one-time problems, such as the amnesties for draft resisters after the Civil War and the Vietnam War, but illegal immigration is a continuous phenomenon. This amnesty is generating expectations of future amnesties, providing incentives for current and future illegal immigration. Those foreign nationals who obeyed U.S. immigration law and are waiting in the lengthy queue to receive their visas must be questioning their own judgment. Furthermore, the current amnesty is already being criticized for the stringency of its provisions, the required long duration of residence and the shortness of the application period. Some of the proposals for reforming the reform would grant amnesty to all illegal aliens who entered the U.S. prior to IRCA's enactment. Finally, as will be shown below, the employer sanction provisions are unlikely to reduce illegal immigration materially.

IRCA demonstrates consciousness of the potentially large impacts on the income transfer system of the current amnesty program. Those granted legal status and their dependents are barred from receiving welfare benefits for the first five years after being legalized. This provision may discourage some otherwise qualified illegal aliens who are now receiving or expect to receive welfare benefits from coming forward, but it is unclear whether this prohibition could withstand a court challenge. It is doubtful that legalized aliens in dire economic conditions who are otherwise eligible could be denied the "equal protection" of the income transfer system, just as the Supreme Court decided that illegal alien children could not be denied "equal access" to the Texas public schools.

Enforcement—The Toothless Tiger

IRCA includes several enforcement provisions, but they are unlikely to have a major impact in slowing illegal immigration.

The Effect of Employer Sanctions. Employer sanctions are intended to be the main enforcement arm of IRCA. The Act makes it illegal for employers to hire workers without legal rights to work in this country. There is no small-firm exemption, a provision common in other labor legislation. The law covers all employers, including households that hire domestics or other workers on a regular basis. A grandfather clause protects employers from penalties for illegal aliens hired prior to the enactment of IRCA.

Employer sanctions are to be enforced by INS, with the assistance of the Department of Labor. Violators are subject to civil penalties (fines) that increase with the number of violations. Criminal penalties (fines and prison terms) can also be imposed when there is a "pattern or practice" of violations.[5] Employers who hire illegal aliens are not subject to penalties as long as they make a good-faith effort to determine the worker's legal status. This defense can be established if the employer examines documents which appear to be genuine or if the worker is referred by the state employment agency, a union hiring hall or a labor contractor who ascertains the worker's legal status. The employer need not verify the authenticity of the documents. Appropriate documents include a U.S. passport, certificate of naturalization, alien registration card, and an INS work authorization permit. Most American workers would not have these documents. They can, however, use a social security card or a birth certificate in conjunction with a driver's license or comparable document. Employers are required to complete and retain a new form, the I-9, on which they record which documents were presented.

The intent of employer sanctions is to reduce the demand for illegal alien labor, but their likely effectiveness is unclear. A survey conducted in 1984 of about 400 employers provides some interesting insights on employer perceptions (Chiswick, 1988, Ch. 5). In this survey, where the respondents were the persons in charge of hiring non-supervisory male workers, about 70 percent of the establishments identified by an apprehended illegal alien and an even higher proportion (about 90 percent) of randomly selected establishments reported that they believed it was a violation of the law to hire an illegal alien. This was two years before IRCA was enacted. This finding suggests that the mere passage of employer sanctions simply puts into law what the employers already believed to be the law, and is thus not likely to influence employer behavior.

A reluctance to convert employers into enforcers of immigration law delayed the enactment of IRCA, and that reluctance persists. In addition, some oppose employer sanctions because it raises the cost of hiring workers, others because it may result in increased discrimination against workers who look or sound foreign. These concerns have been addressed in the legislation and regulations. For example, employers need not go through a costly verification process; it is sufficient that the documents appear to be genuine. Also, IRCA makes it illegal for employers with four or more employees to discriminate against those with legal rights to work; it lowers the firm size limit in the prohibition of discrimination on the basis of national origin and, for the first time, bars private sector employment discrimination against lawful resident aliens.

---

[5] Each physically separate subdivision of a firm that conducts its own hiring constitutes a separate entity under the law. The civil penalty for a first offense is a fine that may range from $250 to $2,000 per illegal alien. The fines are higher for subsequent violations. The criminal penalties for a "pattern or practice" of violations may range up to $3,000 per illegal alien and a prison term that may be up to six months for the entire pattern or practice. The law also prohibits "indemnity bonds," guaranteed payments by workers against any liability that a firm may incur from hiring the worker. The Act provides for hearings before administrative law judges (U.S. House of Representatives, 1986).

Since employers are not legally required to verify the authenticity of the documents shown to them, two types of fraud can be expected. One is the use of counterfeit documents. A market for counterfeit immigration-related documents, especially social security cards, has existed for some time, and this market can be expected to continue to grow. Another is the fraudulent use of valid documents, such as using another person's valid social security card or birth certificate.

Employers are caught between the hammer of costly employment discrimination suits and the anvil of costly verification procedures. They may find that their cost-minimizing solution is to accept at face value whatever documents they are shown.

Enforcement Resources. IRCA provides for an increase in enforcement resources. Curiously, IRCA specifies that the Border Patrol be increased by at least 50 percent. While this will increase apprehensions at the border, the deterrent effect on the illegal population may be minimal if the aliens continue to view the southern border as a revolving door. Although some plans call for increasing the INS investigative unit which does the interior enforcement from the current level of 700 positions, it does not appear that the necessary resources will be appropriated. The Department of Labor's wage and hours inspectors are having their responsibilities expanded to include inspection of the I-9s. However, the Labor Department agents are few in number and there are apparently no plans for expanding the force.

As a result of the minimal change in effective enforcement resources, the probability that employers will be detected in violation is very low. The scarce enforcement resources mean that inspections will be infrequent, except perhaps for certain target or showcase establishments.

The very low expected value of penalties if an establishment is in violation and the ease of establishing an affirmative defense to avoid the detection of violations means that employer sanctions are not likely to have a major impact. The primary impacts will be a relative increase in the hiring costs of labor in low-skilled high-turnover jobs, an increase of the paperwork burden on all employers, and some impact on illegal alien employment in targeted sectors.

An Ignored Enforcement Instrument: Illegal Alien Sanctions. What is perhaps most remarkable about the decade-old debate regarding sanctions as a means of reducing illegal immigration is the nearly exclusive focus on employers, and the absence of a discussion of sanctions on the illegal aliens themselves.

Following the pattern established in previous immigration laws, IRCA does not include "illegal alien sanctions." Other than leaving the country (often only temporarily), there are generally no penalties for illegal aliens who are apprehended, especially for the vast majority who accept voluntary departure. Those who insist on a formal deportation hearing can even delay their departures for extended periods of time, perhaps indefinitely.

Illegal aliens are generally fully aware of their illegal status. Their entry into an illegal status is generally a willful act rather than an inadvertent one. The net earnings from employment in the United States, and hence the attractiveness of U.S. jobs, could be reduced by imposing penalties against apprehended illegal aliens, even those at the border who accept a voluntary departure. The penalties can be in terms of money (fines), time (detention), and administrative sanctions (like denying for a period of time any rights to an immigrant, visitor, student, or temporary worker visa). For many illegal aliens, particularly those who come to the United States for seasonal urban or rural jobs, a detention of two to four months might have a substantial effect on their expected net earnings. To prevent the immigration courts from becoming overly congested, plea bargaining on fixed terms might be used to discourage protracted deportation hearings. Once the certainty of meaningful penalties if apprehended is established, the flow of illegal aliens, particularly across the southern border, would decrease. Perhaps the absence of a serious policy discussion of illegal alien sanctions reflects a general reluctance to enforce immigration law meaningfully.

Conclusions

U.S. immigration policy has been confronted by a dilemma. Low-skilled foreign workers can have positive net benefits on the economic well-being of the native U.S. population. Yet, if they are accorded all of the privileges of legal immigrants they can bring dependent family members and have access to the income transfer and social service systems. In this case they are likely to have a negative impact. One solution is a large guest-worker program, as under the *bracero* program which functioned from 1942 to 1964. These programs seem best suited to seasonal farm work, particularly if the objective is to discourage permanent settlement. Since the end of the *bracero* program, the main answer to the dilemma has been to tolerate illegal employment but discourage the entry of dependents. With the passage of time, however, families are established and the illegal aliens either become "second-class" residents outside the legal system or illegally use transfers and social services. The emergence of a "permanent" illegal alien population was in part encouraged by enforcement policies. Most resources were devoted to enforcement at the border with relatively few resources devoted to interior enforcement. Although the latter is more costly per apprehension, it has a far greater deterrent effect because it imposes more costs on the illegal aliens.

IRCA was intended to change these conditions by "wiping the slate clean" (amnesty), so that the U.S. could start over and close the door to illegal immigration. This was to be done through a combination of employer sanctions and a seasonal farm worker program. Yet, the objective is not likely to be accomplished. Amnesty may increase the supply of low-skilled workers, and among some encourage hopes of future amnesties. Furthermore, millions of aliens remain in an illegal status. The employer sanctions provisions and minimal enforcement resources virtually guarantee that the sanctions will have little deterrent effect on illegal alien employment.

The absence of substantive penalties for apprehended illegal aliens and the continuation of a policy of minimal resources devoted to the interior enforcement of immigration law means that the policy instruments most likely to deter illegal immigration have been ignored. By ignoring these policy instruments during the past few decades the United States has implicitly condoned the establishment of a large resident illegal alien population. This complicity may be the most compelling argument for the 1986 amnesty. By ignoring these policy instruments in the future, and by deluding ourselves by continuing to focus on policies that are likely to have a minimal impact on deterring illegal immigration, the United States can anticipate a resumption in the growth of a resident illegal alien population. Unless more effective policies are implemented, illegal immigration will be a continuing feature of the American labor market and society.

References

Chiswick, Barry R., "The Impact of Immigration on the Level and Distribution of Economic Well-Being." In Chiswick, Barry R., ed., *The Gateway: U.S. Immigration Issues and Policies.* Washington, D.C.: American Enterprise Institute, 1982, pp. 289–313.

Chiswick, Barry, R., "Illegal Aliens in the U.S. Labor Market: Analysis of Occupational Attainment and Earnings," *International Migration Review,* Fall 1984, *18,* 714–32.

Chiswick, Barry R., *Illegal Aliens: Their Employment and Employers.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1988.

Cross, Harry E., and James A. Sandos, *Across the Border: Rural Development in Mexico and Recent Migration to the United States.* Berkeley: University of California Press, 1981.

Cuthbert, Richard W., and Joe B. Stevens, "The Net Economic Incentive for Illegal Mexican Migration: A Case Study," *International Migration Review,* Fall 1981, *58,* 543–550.

Massey, Douglas S., "Do Undocumented Migrants Earn Lower Wages than Legal Immigrants? New Evidence from Mexico," *International Migration Review,* Fall 1987, *20,* 236–74.

Passel, Jeffrey S., and Karen A. Woodrow, "Geographic Distribution of Undocumented Immigrants: Estimates of Undocumented Aliens Counted in the 1980 Census by State," *International Migration Review,* Fall 1984, *18,* 642–71.

Seigel, Jacob S., Jeffrey S. Passel, and J. Gregory Robinson, "Preliminary Report on Existing Studies of the Number of Illegal Residents of the United States." In Select Commission on Immigration and Refugee Policy, *U.S. Immigration Policy and the National Interest,* Appendix E, *Papers on Illegal Migration to the United States.* Washington, D.C.: U.S.G.P.O., 1981, pp. 13–40.

U.S. Department of Justice, *1985 Statistical Yearbook of the Immigration and Naturalization Service.* Washington, D.C.: U.S. G.P.O., 1986.

U.S. House of Representatives, *Immigration Reform and Control Act of 1986: Conference Report,* Report 99-1000. Washington, D.C.: U.S.G.P.O., October 14, 1986.

Wright, Gavin, "The Economic Revolution in the American South," *Journal of Economic Perspectives,* Summer 1987, *1,* 161–178.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 117–131

The Political Economy of Immigration Law: Impact of Simpson-Rodino on the United States and Mexico

Clark W. Reynolds and Robert K. McCleery

About two years ago the United States passed the Immigration Reform and Control Act, also called the "Simpson-Rodino" bill of 1986. The Act called for increased enforcement of migration policy, employer sanctions, and amnesty for those who could prove continuous residence since 1982. Despite considerable discussion and debate prior to the act, the legislation was passed without any comprehensive economic analysis of its potential impact on the United States or its main source of undocumented immigration, Mexico. Notwithstanding the politically charged atmosphere surrounding the bill, the sense among many in Congress and the White House that "we need to restore control over our borders," and the determination by Mexico to stay out of what it regarded as a domestic U.S. legislative matter, the absence of attention to basic economics in a bill so fundamental to the performance of both countries was surprising.

In this paper we shall look at some implications of the recent immigration law for both economies, given their widely differing levels of income and productivity, the challenges each faces to restructure its economy given increased international competitiveness, and the particular problems and opportunities presented by a common border with growing labor market interdependence. By our calculations, the economic opportunity cost of Simpson-Rodino as compared to continuation of the prior status quo will add up to a present value of $110 billion between now and the year 2000.

A general equilibrium approach to the problem of migration policy indicates that the United States has even more to gain than Mexico from greater accessibility to Mexican labor in the next decade than Simpson-Rodino would permit. However, the gains to capital and the majority of higher skilled workers would be partially offset by slower earnings growth for the lowest 10 percent of its wage earners. Mexico, on the

■ *Clark W. Reynolds is Professor of Economics, Food Research Institute, Stanford University, Stanford, California; Robert K. McCleery is a doctoral candidate, Department of Economics, Stanford University, Stanford, California.*

other hand, would see gains from a relaxation of Simpson-Rodino for actual and potential migrants (the majority of its work force), but somewhat slower increases in profits and wages of upper income groups. Interestingly, the net gains to each country from increased migration are shared across the border by different groups: U.S. capitalists and higher-skilled workers together with Mexican migrants and lower-skilled workers. By the same token, the losses from increased migration lead to peculiar cross-border alliances of Mexican capitalists and higher-skilled workers together with U.S. lower-skilled workers.

The results indicate that the political economy of migration policy is truly transnational. Labor market interdependence links economic policy between countries, whether one likes it or not. Benefits and costs overlap the border in terms of sectors, regions, and income groups. Trade and industrial policies in one country influence the labor market of the other, as do patterns of investment and technological change.

The fact of increasing labor market interdependence between the United States and Mexico, owing to proximity, demographic trends, growing wage differentials, and rising social aspirations, will demand much greater cooperation in policy-making than occurred in the case of Simpson-Rodino. In fact, Simpson-Rodino illustrates the important role that labor mobility may play in the convergence of income and productivity between rich and poor countries. It shows how migration policy may distort or delay that process of convergence, with negative implications for both societies. Such an approach indicates how trade, capital, and technology flows may complement or substitute for migration in a period of ever-increasing international exchange, depending on the degree of international cooperation in policy formation.[1]

In this regard, "loss of control of the borders" may be a consequence of restoring control over such policy objectives as competitiveness and productivity growth, by raising decision-making to the international level. The goals of full employment, improved wages, hours, and working conditions, and investment profitability in the United States and Mexico are in fact related. Labor market interdependence calls for policy coordination in the two countries, if both are to gain from their unique and potentially productive relationship.

Alternative Approaches to Migration

The simplest analysis of migration, the *gravity model*, deals with a flow between markets in response to expected earnings differentials. Movement tends to take place at the margin, so that those who leave the low wage market are least employable, and those positions in the high wage market filled by immigrants are least attractive. Under the gravity model, migration raises productivity of labor in the sending region

---

[1] The *factor price equalization theorem* states that under certain strong assumptions free trade will bring about wage equalization between countries. However, the U.S.-Mexico case illustrates that where a large percentage of the work force of the low-wage country is employed in the production of nontradeable goods and services, and where a growing percentage of employment and output in the high-wage country are also in the nontradeable area, and where labor markets are subject to segmentation favoring wages in the tradeables sector, even a dismantling of current trade barriers would not bring about wage convergence. Migration in this case is essential to accomplish a reduction in inequality between the two countries.

(by raising the capital/labor ratio) while raising output, employment, and the productivity of capital in the receiving region. The simple gravity model sees increased exchange between markets as providing gains to society at large, although they may not be Pareto optimal, since low wage workers in the high-wage country might be disadvantaged, along with capital in the low-wage country.

In fact, however, labor movement involves challenges which include the cost and uncertainty of job search, travel, border crossing,. resettlement, and overcoming language and cultural barriers. Hence, those who do migrate tend to have higher education, skills, ambition, and more savings than their poorest compatriots. Their loss imposes a greater opportunity cost to the nation than would be the case if the least productive workers were to migrate. But average productivity of the sending economy will still increase as a result of out-migration.

The *brain drain* approach to migration analysis stresses that a significant component of human capital (and sometimes financial capital) is transferred with the worker, tending to lower labor productivity in the sending region while raising it in the receiving region. Often such migration is motivated by social, cultural and professional reasons as well as by wage differentials. The effect of migration under such circumstances is the opposite from that of the gravity model: Here, factor movement has the perverse effect of widening inequalities between sending and receiving regions. It is often associated with inadequate levels of development and lack of opportunity for educated labor in sending regions.

In the case of Mexico, both approaches to migration have relevance. The gravity model is more applicable to the years of stability and growth before the 1982 debt crisis, while the distortions and uncertainties of the crisis have given rise to another dimension of migration to which the brain drain model may apply more often. Recent devaluation of the peso, stagflation, and severe underemployment have tended to increase the cost of migration, making it harder for the poor, while widening the earnings gap for higher skilled labor and professionals.

If employment opportunities in Mexico languish for skilled labor and professionals as educational levels improve in that country, and wage differentials widen for upper income cohorts, the brain drain model may increase in relevance. At present, however, the gravity process continues to dominate. Many immigrants are still from the lower income groups, rural areas, and small towns, with education and earnings levels far below the national average, though they are above the least educated and poorest cohorts.[2] For instance, Taylor (1987) studies selection bias, human capital, and undocumented migration and concludes:

> There is no evidence that people who migrate illegally to the United States are above average contributors to household income, either as workers in Mexico or as Mexico-U.S. migrants. However, other things being equal, villagers who are

---

[2] Current monitoring of undocumented migrants by El Colegio de la Frontera Norte (1986–88) indicates that the increasing costs of migration since 1982 have changed the profile of migrants. The result is a trend toward urban workers who are better educated. According to these findings the average education of the current undocumented migrant exceeds the national average for Mexican labor while only one third proceed from rural areas and small towns [conversations with Jorge Bustamante, May 1988].

in the best position to contribute to household income as workers in Mexico are positively selected not to migrate to the United States.

Decisions to migrate tend to be based on the individual's comparative advantage, with education and work experience at home raising potential earnings in Mexico relative to the United States, while previous experience abroad increases the relative earnings from return migration. In the case of those who do migrate with higher than average skills and education, the dampening effect on domestic productivity is offset by remittances in the short run and experience and training brought back to Mexico in the longer run.

Macroeconomic Policy Distortions and Other Short-Run Impacts on Mexico-U.S. Migration

A perception has been growing in the United States that migration from Mexico is a "problem" that demands a "solution," but this feeling has tended to make migration a scapegoat for divergent macroeconomic policies. By the late 1970s, evidence of rapidly increased immigration (including refugees, legal, and undocumented workers) had become overwhelming. While Simpson-Rodino was a response to global perceptions and concerns, calls for "the restoration of control over our border" were bound to have a disproportionate impact on Mexico since it was the largest and most visible source of undocumented labor. The issue had been exacerbated by macroeconomic policies in both countries and the interaction between them.

In the United States, measures taken to reduce inflation had by the beginning of the 1980s raised real interest rates, slowed the economy, and increased unemployment. Concurrent structural changes in the job market led to a deceleration and even decline in real wages along with a rising share of low-wage/low-skill jobs. Labor power was weakening and conditions worsening at the very time that new immigrants were becoming more visible. Not surprisingly, immigrant labor was blamed for the problems of domestic workers.

In part, conditions in the U.S. labor market reflected growing international competition which was exacerbated by the overvalued dollar[3] that resulted from tight monetary policy in the fight against inflation. A lot of U.S. manufacturing investment flowed offshore in those years as tax incentives, wage differentials, and the dollar's overvaluation led to a shift of production and employment to overseas suppliers. These elements alone would have explained much of the slackness in the U.S. labor market—which peaked in 1982 at 9.5 percent unemployment—quite independent of immigration in general, much less from Mexico.

[3] While the magnitude of the impact of Simpson-Rodino is not yet clear, any given degree of enforcement can be studied with our approach. Over- and under-valuation of currencies as used in this article are measured relative to purchasing power parity.

Table 1
Recent U.S.-Mexico wage divergence

	U.S. Minimum Wage	*Mex. Minimum Wage*[a]	*Gap (U.S. dollars)*
1978	$2.65	$.81	$1.84
1980	3.10	.97	2.13
1982	3.35	.27	3.08
1984	3.35	.56	2.79
1986	3.35	.55	2.80

[a] Mexican minimum wage in Tijuana (per hour of eight hour day) in current dollars at free exchange rate; Tijuana and other border regions are above the national average. Source. IMF, *International Financial Statistics*; Comision Nacional de Salarios Minimos. Minimum wages in Mexico either directly cover or are linked to market wages in many sectors and occupations.

In Mexico's case, an oil-led boom caused the economy to grow by over 25 percent during the period 1978–81. This led to growing overvaluation of the peso and temporary tightening of the job market. However, the government continued the program of import-substituting industrialization, a strategy that proved to be highly capital- and import-intensive, with an output elasticity of employment of only 0.4. Meanwhile, population growth and rising female labor force participation rates were causing an increase in labor supply. Peso overvaluation discouraged labor-absorbing investment in agriculture. Growth of the government sector as a major employer did not favor those most likely to migrate abroad. Thus, despite the oil revenues, pressures were building toward increased levels of migration.

In 1982 the oil bubble burst and Mexico found itself with an increasingly costly debt (as real interest rates soared) and weakening export revenues. For the first time since the great Depression, the economy declined and the shock was felt by all sections, particularly labor. Considerable flexibility in the labor market caused a price rather than quantity adjustment, and from 1982 to 1986 real wages fell by over 40 percent. The government responded to the debt crisis with a program of fiscal stringency, initial import controls, and massive devaluation of peso, followed by a program of trade liberalization, reduced subsidization of consumables, and mini-devaluations that shifted the real exchange rate from about 20 percent overvaluation before 1982 to 20 percent or more undervaluation thereafter. The result, as shown in Table 1, was a sudden increase in the wage gap between the United States and Mexico at the very time when the labor market was particularly slack in both countries, especially in Mexico.

The shift from over- to under-valuation of the peso reflects a swing from heavy capital inflows during the period of oil-based borrowing to significant net outflows for debt service after 1982, as the country shifted from a strong deficit to surplus trade balance. This swing was important, because decisions to migrate are in fact heavily influenced by the purchasing power of the respective currencies. To the extent that debt resolution will permit the reestablishment of reasonably stable exchange rates (as measured by purchasing power parity) and restored growth in Mexico, one would expect debt resolution to lead to a considerable lessening of migratory pressures.

Modeling Medium and Long-Term Labor Market Interdependence

The goal of this analysis is to estimate the efficiency gains from migration. These gains are customarily ignored in studies of migration that deal with the optimal imposition and enforcement of barriers against immigration per se and employment of undocumented aliens. By contrast, we believe that economic benefits that would result from unconstrained migration should be calculated first, and then these benefits can be compared with the social, political, and distributional problems associated with migration. The fact that the gains from migration are unevenly distributed between sending and receiving country and among social groups, regions, and sectors of each country makes this a difficult problem of political economy.

The Mexico-U.S. case allows one to focus on such economic aspects of undocumented migration because individuals migrate in order to work rather than working because they have migrated. In the Mexican case, the sociological and political influences which are so important in explaining much of recent migration to the United States are secondary to economic factors. In this case migration can be dealt with as an example of labor market interdependence, just as trade is a reflection of product market interdependence. Attempts to impose barriers between markets will be met by extralegal flows in response to the differentials in wages (or prices) created by those barriers. To measure the impact of such barriers or their removal, it is necessary to model the wage and employment determination process in both sending and receiving countries, with migration explicitly included. Over time, demographic and policy changes combine with conventional sources of growth (savings rates, physical and human capital accumulation, technological progress) to determine the time path of labor productivity, wages, and migratory response.

The authors have formulated a model to address the type of questions raised above (Reynolds and McCleery, 1985; McCleery, 1988). This time path of migration, wages, and employment is presented in terms of a range of hypothetical policy approaches ranging from free migration to total blocking of the border, recognizing that neither extreme is realistic. The model is designed to answer the following questions: What time path of migration would have occurred in the absence of Simpson-Rodino under the status quo; with an open border; with a closed border? How are the results of Simpson-Rodino likely to differ from the status quo (under the strong assumption that the provisions of Simpson-Rodino are fully enforced)?[4]

[4] The results generated by a dynamic, general equilibrium model such as we have employed may be compared with a simple static equilibrium approach, as the following exercise illustrates. Suppose that wages reflect the marginal productivities of workers in both countries and that migrant and domestic labor are perfect substitutes. Then one can obtain a rough estimate of the welfare gain to individual migrants by measuring the difference in real wages, holding the quantity of other factors constant. If the relevant wage were $9000 in the United States and $2000 in Mexico, representing a wage gap of $7000 per year, a policy

Before looking at the implications of Simpson-Rodino for the two economies, the model examines what would have happened in the absence of the legislation between 1985 and the year 2000 (the "status quo scenario"). Taking savings rates, technological progress, and tastes as exogenous, the simulation yields output, employment, wage, migration, and trade estimates by country and sector (high-wage tradeable and low-wage nontradeable) within each country for each year between 1986 and 2000.[5]

Output growth in the U.S. varies around a trend rate of 3 percent and the Mexican growth rate recovers to 5 percent per annum by the mid-1990s. Wage growth is concentrated in the high-wage sectors of both countries, as the inflow of migrants raises output and employment, while lowering consumer prices. This raises the real incomes of all but the 10 percent of the labor force in direct competition with migrant labor from Mexico. These projections show a rise in Mexican undocumented migration levels (man years of labor) from an estimated 2.5 million in 1982 to 3.1 million in 1985, peaking at over 4 million in 1996 and falling back to about 4 million in the year 2000, with a continued decline thereafter as the economies converge. The 1985 figure represents less than 3 percent of the U.S. labor force, but may account for more than 20 percent of selected occupations in the low-wage sector.

---

that reduced migration by 500,000 unskilled workers would have an opportunity cost of about $3.5 billion. (Note that the estimated reduction in migration must be obtained independently here, as compared with the general equilibrium approach.) This opportunity cost can be improved by taking into account the impact of the reduction in migration on the wage level of the sending and receiving countries. Given the initial wage figures above, and assuming that the elasticity of demand for labor in each country is near unity, then the 2 percent increase in the number of unskilled workers in Mexico would reduce wages by $40 while the 3 percent reduction in the number of unskilled workers in the United States would raise wages by $270, increasing the total wage gap to $7310. This gives a total opportunity cost of almost $3.7 billion.

Additional complexities that could be treated in a partial equilibrium approach but usually aren't are the changes in the return to capital and skilled labor in each country. For example if there is a positive elasticity of substitution between the two skill levels, then the gap between skilled and unskilled labor will widen in Mexico and narrow in the United States.

For the same cutback in migration, our general equilibrium model gives a $5 billion opportunity cost. Normally the various feedback and adjustment mechanisms in a general equilibrium model mitigate the effects of partial equilibrium analysis, but in this case our opportunity cost differential is much greater in the general equilibrium model because of the negative impact on U.S. capital of the loss of unskilled labor (which was not considered in the partial equilibrium exercise above) and the resultant economic adjustments.

The static nature of the partial equilibrium calculations produces an underestimate of the welfare cost of reducing migration because in the absence of more migration than Simpson-Rodino would accomodate, as a result of current demographics and capital constraints, wage differentials would be expected to increase during the next two decades. While the Mexican wage of unskilled labor is currently about twenty percent of the United States level, we estimate that in the absence of migration by the year 2000 it will have fallen to 14 percent. These partial equilibrium calculations do not take into account any macroeconomic effects, including responses of other factors' supplies to the changes in factor returns. However, the general equilibrium model does take such elements into account along with the dynamics of wage adjustment. The general equilibrium model also takes into account the fact that demand and supply of each country's output will change, altering internal and external relative prices. When we incorporate all of these elements into our analysis, the annual estimate of the welfare loss of reducing migration through Simpson-Rodino (compared with the "status quo" scenario) grows to over $7 billion by the year 2000.

[5] See the Appendix for an overview of the methodology of this model.

The status quo and the Simpson-Rodino bill can be thought of as two points on a continuum of migration policies ranging from "free migration" to the complete closing of the border. It should come as no surprise that there is a substantial economic benefit to labor movement from the low-wage, labor-intensive sectors in Mexico to the higher-wage, relatively capital-intensive U.S. economy under any scenario in which migration is allowed to take place. With free migration (notwithstanding the clear social and political costs) these efficiency gains (relative to a closed or autarchic labor market) amount to about 2 percent of the combined GNP of the two countries per year, with the discounted sum (in 1986 dollars) of all future returns amounting to nearly $1 trillion.

Of course, we are not by any means advocating a free migration policy by presenting such hypothetical figures, but rather are attempting to illustrate that the appropriate migration policy must take into consideration significant opportunity costs.

In addition to a stream of benefits, each scenario gives a different path of migration, though all have the same basic form. All peak around the year 1996, when the Mexican population growth slowdown of the mid-1970s is translated into a decline in labor-force growth rates. The free migration level in the year 2000 is over six million migrants, or some 15 percent of the Mexican labor force, which is some 50 percent above the status quo level for that year and almost twice the current level.[6]

Thus each scenario can be evaluated in terms of the stream of costs (proxied by average migration level) and benefits it provides. These benefits are calculated as the discounted sum of the compensating variations across all social groups in both countries. In other words, they reflect the total amount of income needed to compensate all individuals for the decline (or increase) in their utility between the two scenarios. Consumer demand theory tells us that a point with at least as much benefit and less migration would be preferred. But since the status quo would lead to both more migration and more benefits than Simpson-Rodino, neither scenario can be said to dominate unless the tradeoff between the costs and benefits of migration are more fully specified. For the moment we will postpone questions of distribution, both within and between countries, and focus on characterizing these overall pictures.

The old status quo would have resulted in a relatively open border, with two-thirds of the possible migration and two-thirds of the benefits of migration. In light of Simpson-Rodino, this path of migration, rising to over 4 million man-years of labor, might be seen as unacceptable by the United States in 1986 for reasons we shall examine later, though of course no such projection figured in the legislation process.

The method of reducing migration that was settled on was the Simpson-Rodino bill, which included provisions for increased enforcement, employer sanctions, amnesty, and a three-year special visa program for agricultural workers. Our projections indicate that this legislation—if strongly enforced—would be moderately successful in

---

[6] It should be stressed that this is a two-country model; if the United States were to open its borders to all, immigration would originate from many other countries besides Mexico, reducing the number of Mexican migrants while increasing the total inflow in view of greater wage gaps elsewhere.

Table 2
The cumulative impact of Simpson-Rodino on social groups in the U.S. and Mexico from 1986–2000

Social group	*United States*	*Mexico*
	(all numbers are billions of U.S. dollars, discounted to 1986)	
High wage workers	−50	+5
Low wage workers	+50	−12
Capitalists	−100	+5
Migrants	—	−8
Total	−100	−10

reducing migration,[7] dropping the year 2000 migration level well below the four million mark of the status quo scenario, including at least half a million undocumented migrants legalized under the bill.[8] Yet this reduction in migration will come at a substantial cost by the year 2000—over $100 billion 1986 dollars, about 3 percent of the combined 1986 production of the two countries and over one-eighth of the potential benefits of migration. These costs will be borne primarily by the U.S. economy, which is $100 billion worse off, while Mexico loses over $10 billion. Table 2 presents an accounting of these costs.

The costs of Simpson-Rodino include both the production and consumption reductions resulting from the international misallocation of resources and the opportunity cost of resources used to enforce and evade the new law. Thus, the law introduces additional inefficiency at any given level of migration, in addition to the cost of lower migration levels. Higher INS budgets, in addition to their explicit cost, will raise the cost of border crossing and the average amount of resources expended in the crossing (like fees to "coyotes," those who smuggle migrants across the border, and the purchase of false documents). More detentions and deportations will increase the time workers are out of both countries' labor markets. Employer sanctions will act to increase the spread between the marginal product of undocumented migrants and their wage, to the extent that its incidence falls on labor. This spread distorts the pattern of resource use in the United States. Indeed, the differential risk of hiring undocumented workers between small and large firms could influence firm size; and if these considerations worked contrary to increasing returns in production, another source of inefficiency is introduced.[9]

---

[7] The penalties and increased INS activity levels associated with Simpson-Rodino are translated into costs borne by migrants and employers, driving a wedge between the marginal product of the migrant and his wage. The modeling assumption is that the new, higher level of INS activity is maintained throughout the period. In fact the level of INS activity appears to be related to the level of U.S. unemployment.

[8] If the first year of the bill had an estimated thirty percent negative impact on the gross flow of Mexican immigrants, this would translate to about a ten percent reduction in the stock of undocumented labor, which is consistent with the figure used in our projections. The Immigration and Naturalization Service reported that U.S.-Mexico border apprehensions for the fiscal year ending September 1987 were down 30% from the previous fiscal year. For the first quarter fiscal 1988 (October-December 1987) apprehensions were down 36% from the same period in the prior year [New York Times, January 15, 1988]. Over time, however, the effect of the law will reflect the degree of enforcement of employer sanctions and apprehensions, both of which depend upon appropriations. It is possible that the stock could decline by well over ten percent in time.

To do cost-benefit analysis even at this level of abstraction, one needs to quantify the costs of high migration levels. Supposing for the moment that Simpson-Rodino is a rational action, what benefits exist to offset the costs that have been documented?

One justification for the Simpson-Rodino bill has been that migrants tend to compete for jobs with youth, minorities, and women, groups that already have high unemployment rates and relatively low wage levels. In this light, further increases in migration would increase the hardships of these already disadvantaged groups, while the migration reductions in Simpson-Rodino will have a positive effect on income distribution. This reasoning is correct as far as it goes. In our model, the hundreds of thousands of migrants deterred from the low-wage sector in the U.S. make each of the over 8.5 million workers in the sector better off by about $650 in 1986. The corresponding negative effect on the incomes of other Americans is much smaller, being the indirect effect on real income as prices rise to equate supply and demand for the much lower output of the low-wage sector. A second indirect effect comes from the redirection of investment toward this sector, as higher wage costs lead to more mechanization. In welfare terms, these effects sum to less than $50 per person annually, but apply to all of the nearly 100 million workers in the economy, nearly canceling the benefits to low-wage workers.

Owners of capital are hurt by the exodus of labor, as the return to capital is inversely related to the capital/labor ratio. The total loss to owners of capital is nearly $10 billion, leaving a net loss to the U.S. of about $8 billion. As mentioned above, the loss to capitalists would be greater if structural rigidities led to bankrupcies or idle capacity, but would be less (and there would be a corresponding fall in the gains to low-wage labor) if capital were free to move offshore.

Thus, the Simpson-Rodino bill will have the effect of reducing migration, and that reduction will make income distribution more equal in the United States. However, it seems that Simpson-Rodino is a very costly way to achieve improvement in income levels of the American poor, if that is indeed its intent. One should be able to use the existing mechanisms of taxes and transfers to reach the goal in a way less distortionary and costly for the economy as a whole. Further, enforcement of Simpson-Rodino will lower incomes and worsen distribution in Mexico, a country with far poorer poor and a more serious income distribution problem.

The Mexicans who might be crowded back into Mexico by Simpson-Rodino will exacerbate two problems in the Mexican economy. As mentioned before, the loss of remittances will accentuate the foreign exchange crisis, and the returning migrants will swell the ranks of the underemployed. The former results in a $500 million reduction in intermediate imports in 1986, while the latter reduces the welfare of low-wage Mexicans by almost $90 a year.

---

[9] Not all of these effects are incorporated into the formal migration model used in this paper, but they are presented as further evidence of the manner in which Simpson-Rodino will impact the U.S. economy.

The relatively well off in Mexico, on the other hand, will benefit from this legislation. Skilled labor gains from the reduction in unskilled wage rates through the fall in prices, and capitalists gain from the increase in the labor force in the same manner U.S. capitalists are hurt. Yet these gains are mitigated by the fall in intermediate imports and the resultant contractionary effect on the industrial sector. Thus, the net effect for the country as a whole is a loss in utility, despite the rise in total production, as the average productivity of those employed in Mexico falls. The catch is that per capita production declines by about 4 percent as the returning migrants drive down average labor productivity.

The bottom line is that the only economic justification for the passage of the Simpson-Rodino bill is its effect on income distribution in the United States, and that benefit is only achieved at high cost to Mexico and the rest of the U.S. workers. It is not difficult to see how significant political support for this bill was motivated. Low income U.S. workers were significantly harmed by the current and projected levels of migration under the old migration regimen. The unemployed are easy to convince that migration is the root of their problem, and labor unions resist the inflow of difficult to organize and potentially competitive labor. Middle and upper income voters in the United States who are not in competition with migrant labor are the ones who benefit from a relatively open migration policy. Yet these benefits are indirect, through a reduction in prices, and are spread rather thinly among a prosperous group. If this group perceives an influx of those with different religious beliefs, culture, and lifestyles to be a threat, they might value the preservation of their political power, assets, and way of life in general more than a small increase in purchasing power. Capitalists have a similar problem in organizing opposition to the bill. The vast majority of undocumented workers are employed by small firms with less than 50 employees, and thus have little political clout or resources to work with. The area in which large firms and co-ops employ migrants, agribusiness, has a three-year exemption from the restrictions of Simpson-Rodino. An additional reason for smaller firms not to fight the bill might have been the potential alienation of one's domestic work force and customer base.

The position of Hispanic organizations on the bill remains ambiguous, reflecting contradictory economic, social, and political elements. On the economic side, many Hispanic labor leaders saw their constituents' interests threatened by continuation of the status quo, in which undocumented workers were employed at marginal wages and in substandard conditions. However, in terms of the scope for increased organization of workers and political impact, reduction in further immigration would place limits on the growth of Hispanic power. On the other hand, the bill's amnesty program could increase their political power by raising the number of qualified voters. For many the main grounds for opposition to the measure was the potential for discrimination against legal Hispanics in the work place, and the bill included a number of provisions to address this concern.

Conclusions

In cases where markets are contiguous and complementary, powerful economic forces exist drawing them together. This is the case with the U.S.-Mexico labor market. Under such circumstances barriers between the markets will introduce serious distortions in efficiency, equity, and justice. America has seen this in other instances, such as Prohibition, where the barriers were not only illegally surmounted but led to the growth of institutions that were threatening to society. While the United States clearly needs to have a voice in who can enter the country, this should not preclude an orderly and mutually beneficial approach to the particular complementarities that exist between the United States and Mexico in terms of their age and employment structure, skills, demand patterns, and resource endowments.

In every scenario we examined, even with passage (and enforcement) of the Simpson-Rodino bill, migration increases between 1986 and 1996. It is not only the "*push*" *of labor force growth in Mexico but the* "*pull*" *of demand in the United* States that causes the barriers to be surmounted. By the early 1990s Mexico's own growth with restructuring in a more labor-using direction will cause employment growth to catch up with supply and increase wages and productivity. Meanwhile it is no secret that the U.S. population is aging. In addition to the well-documented effects that this will have on the social security system and other aspects of the "dependency burden," there will be an impact on the structure of the national work force and the pattern of consumption that will work at cross-purposes. There will be a surfeit of those pursuing white collar and supervisory positions and too few prepared to fill low-skilled entry-level jobs.

Today, these fundamental complementarities between the U.S. and Mexican labor forces may be obscured by macroeconomic conditions, but we predict there will come a time when the unique labor market relationship shared by the two countries will be recognized by the United States to be as important to its own interests as it is to Mexico. By the same token, a comprehensive examination of Mexico's realistic development options, especially during the critical decade of transition that lies ahead, must include provision for a more effective use of its labor resources than Simpson-Rodino would allow. Indeed, if the existing immigration law is unaltered, its employer sanctions will drive down wages of undocumented Mexican workers and distort employment away from larger manufacturing establishments toward less visible and more exploitative establishments (Hill and Pearce, 1987). The effect will be to skew employment, productivity, and output among sectors and regions of the United States, so that those areas, employers, and jobs that are most able to violate the law will be favored relative to those who comply with it. This is in the interest of neither country.

However, alternative policy approaches that go beyond Simpson-Rodino must come to grips with the adverse impact of Mexican migration on lower-skilled workers in the United States. Owing to the declining marginal benefits of movement toward free migration combined with growing social and economic costs to affected groups, some intermediate policy approach between Simpson-Rodino and open borders would be optimal. A more open migration program could provide benefits to both countries that would be easily sufficient to cover the cost of compensating disadvantaged groups. Such a policy mix might include the following steps.

First, a reduction in bilateral trade and investment restrictions so as to permit both countries to develop along the lines of dynamic comparative advantage. This would permit more efficient use of Mexico's abundant labor, as well as capital and technology in both countries, speeding up their adjustment to a changing international economy.

Second, continuation of favorable debt initiatives such as debt-equity swap arrangements that reduce Mexico's internal and external debt service burden while permitting an increase of investment in those industries favored by the objectives in the first step. These arrangements have been particularly useful in channeling Mexican capital that would otherwise leave the country back into productive domestic investments, especially in labor-using export activities.

Third, the pursuit of innovative rural and regional development programs in Mexico that would include education, infrastructure, and the diffusion of industry and new technologies to absorb labor at ever increasing wages and skills in regions of severe underemployment. U.S. participation in such programs, apart from conventional profit considerations, would be consistent with the raising of low-wage labor productivity in both countries, while reducing migratory pressures.

Fourth, programs that attract emigrant remittances and other financial resources into the above programs could be jointly developed through voluntary swap schemes as have been successfully employed in the European case.

Fifth, a more efficient program of migration regulation than either Simpson-Rodino or the previous status quo is feasible if both countries were to cooperate to their respective benefits while protecting the rights of workers. Under such a mutually regulated temporary visa program, a tax on the windfall benefits to employers and migrant workers could be used to finance retraining and adjustment programs in the United States as well as to supplement support for the above-mentioned Mexican labor-using investment programs. Even if such a tax proved to be unworkable in practice, the use of general revenues for such programs in both countries would be justifiable, given the potential benefits of increased migration under these policies (Reynolds and McCleery, 1985).

Appendix
A Methodology Overview

This model of undocumented migration from Mexico is characterized by temporariness, economic motivation, high participation ratios, and low unemployment. A more general model of legal and undocumented migration from all sources and motivations would obscure the particular nature of this very important and predominant type of migration. For instance, Mexican undocumented migrants have been shown to contribute much more in taxes than they receive in services and transfer payments (Community Research Associates, San Diego, 1980; Undocumented Workers Policy Research project, Texas, 1984).

This appendix is intended to give the reader some insights into the structure and assumptions of the model that was used to generate the results presented here. The interested reader is directed to the McCleery dissertation draft for a complete exposition of the model. The authors took as a point of departure the traditional two-country, three factor (skilled labor, unskilled labor, and capital) trade model and adjusted it to fit the stylized facts of the U.S.-Mexico situation. There are two goods produced in each country via constant returns to scale CES production functions. Labor is sector-specific in the short run while new investment is allocated such that the return to capital is equalized within a country. Skilled labor in each country produces manufactures while output of the unskilled sector produces primary products and most services, especially on the Mexican side. Mexican exports to the United States are treated as consumption goods, and U.S. exports to Mexico as intermediants goods used in the production of manufactures.

The demand specification used is the Stone-Geary linear expenditure system, in which demands are proportional to income above a subsistence level. In the United States demand growth through increases in income favors the low wage sector, since the service sector which employs a growing fraction of undocumented Mexican migrants has a relatively high income elasticity of demand. In Mexico, the demand for manufactures (the skilled sector output) rises with income gains, while demand for agriculture, construction, and services produced in the unskilled sector increases with population growth.

Each economy equilibrates in the short run so as to equalize the marginal product of capital between sectors, and prices adjust to eliminate excess demands. Additionally, migration takes place between the low wage sectors of the two economies until the marginal migrant is indifferent between migrating or not migrating taking into consideration a number of factors, including most importantly the costs borne by the migrant stemming from the activities of the INS. Enforcement of immigration restrictions in the United States drives a wedge between the marginal product of the migrant and the wage he receives. And thus the law introduces a source of inefficiency in the larger economic system which appears as opportunity cost in the model.

The severe balance of payments constraint Mexico faces as a result of debt service payments and oil price fluctuations affects the pattern of trade in this model. In addition, Mexico's need to import intermediate and capital goods from the United States motivates its exports. The model makes allowances for the fact that the Simpson-Rodino bill or any other measure that reduces migration or migrant earnings will adversely affect Mexico's balance of payments through a reduction in migration remittances.

The dynamics of the model are influenced by assumptions about labor growth in each country, the degree of labor movement between sector within a country, and savings and investment behavior. Other important exogenous variables include interest rates, which are assumed to stay constant at 8 percent, and world oil prices, which rise smoothly to $30/barrel by the year 2000. Fluctuations around these trends affect the annual projections but do not change the broad scope of the model's results.

References

Anderson, Annelise, *Illegal Aliens and Employer Sanctions: Solving the Wrong Problem.* Stanford, California: The Hoover Institution on War, Revolution and Peace at Stanford University, April 1986.

Harris, J. R., and Todaro, "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review,* March 1970, *70,* 126–142.

Hill, John K., and James E. Pearce, "Enforcing Sanctions against Employers of Illegal Aliens," *Economic Review,* Federal Reserve Bank of Dallas, May 1987, 1–15.

Kindleberger, C. P., *et al., Migration, Growth, and Development.* Paris: OCED, 1978.

McCleery, Robert K., "Modeling U.S.-Mexican Economic Linkages: A Model of Migration, Trade, and Capital Flows," unpublished Ph.D. dissertation, Department of Economics, Stanford University, 1988.

Ohlin, Bertil, *Interregional and International Trade.* Cambridge, MA: Harvard University Press, 1983.

Reynolds, Clark W., "Labor Market Projections for the United States and Mexico and Current Migration Controversies," *Food Research Institute Studies,* 1979, *XVII,* No. 2, 121–155.

Reynolds, Clark W., and Robert K. McCleery, "Modeling U.S.-Mexico Economic Linkages," *AEA Papers and Proceedings,* May 1985, *75,* No. 2, 217–22.

Stolper, Wolfgang F., and Paul A. Samuelson, "Protection and Real Wages," *Review of Economic Studies,* November 1941, *XI.*

Taylor, J. Edward, "Undocumented Mexico-U.S. Migration and the Returns to Households in Rural Mexico," *American Journal of Agricultural Economics,* August 1987, 626–38.

Velasco, ALfredo, Joseph Nalven, and Craig Frederickson, *Undocumented Immigrants: Their Impact on the County of San Diego.* San Diego, California: Community Research Associates, May 1980.

Weintraub, Sidney, Gilberto Cardenas, et al., *The Use of Public Services Undocumented Aliens in Texas.* Austin, Texas: The Board of Regents of the University of Texas, 1984.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 133–145

Minimum Wage Laws: Are They Overrated?

Charles Brown

Since the Fair Labor Standards Act was passed in 1938, the minimum wage has averaged a bit less than half of average hourly earnings. Typically, the minimum has been fixed in nominal terms at about 55 percent of the current average wage. Rising wages in the rest of the economy gradually erode the real minimum wage and eventually the process is repeated. Sometimes the increase in the nominal minimum is accompanied by a (permanent) reduction in those businesses exempted from the law. Proposals for a lower minimum wage for young workers or to index the minimum wage to the average wage have been debated but not adopted.

The minimum wage was last increased (to $3.35) in 1981. Since then, average hourly wages have risen by 25 percent, eliminating any momentum for a lower minimum wage for youth but reviving proposals for another increase in the nominal minimum.

Economists are fascinated by the minimum wage because we have strong theoretical predictions about the direction of the effect of the minimum wage on employment, and fairly widely accepted methods to estimate the size of that impact. (It probably didn't hurt that the earliest (and later) estimates lined up well with the predictions.) But that is not the only reason for our fascination with the topic. Depending on whether we include those who say they earn *less* than the minimum wage, "minimum wage workers" account for 6 to 12 percent of those employed (Gilroy, 1981, p. 157) and less than 5 percent of wage and salary income. The minimum wage gives economists a relatively rare opportunity to study the impact of an exogenous increase in the price of labor, and thus a chance to refine our thinking about labor markets more generally.

■ *Charles Brown is Professor of Economics and Research Scientist, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts.*

I think there is a reasonably high degree of agreement in studies of the impact of the minimum wage on employment, unemployment, and the distribution of income. For example, Eccles and Freeman (1982) pointed out the striking agreement between research sponsored by the "liberal" Minimum Wage Study Commission and the "conservative" American Enterprise Institute (their characterizations) on the minimum wage. I will also point to a few areas where further work could refine our understanding of labor markets, and perhaps even change the prevailing view of the impact of the minimum wage.

The Textbook Minimum Wage Theory

An introductory textbook without a discussion of minimum wage laws might not be like a day without sunshine, but would certainly rank with a morning without caffeine. With relatively few refinements, that simple model—in which a minimum wage above the free-market equilibrium wage for "unskilled" labor moves the economy back along the demand curve for such labor, reducing employment—really does summarize the heart of the matter.[1] Because we're moving *along* the demand curve, the proportional change in employment in response to a change in the minimum wage is simply equal to the elasticity of demand for unskilled labor times the proportional wage increase. Earnings of unskilled workers as a group rise if demand is inelastic, and fall if it is elastic. But it's important to be careful about a few details in the simple textbook model.

As always, the "prediction" of reduced employment is an other-things-equal claim; one who reads even a tiny sample of public discussion is likely to be amazed how often this point is missed. During the 1977 Congressional hearings on the minimum wage, for example, much was made of the fact that, in the year following previous increases in the minimum wage, employment had risen and unemployment had fallen or remained constant (Subcommittee on Labor Standards, 1977, p. 34, 51). Ten years later, matters were even more confused in testimony before the Subcommittee on Labor Standards (1987, p. 9): "...the arguments that were put forth [in 1977] were, 'There will be a loss of jobs. People will go out of business. It is an unfair imposition.' Again, history and experience point out clearly that there is no basis for that argument. More than 11 million new jobs have been created since 1981 and productivity is up 6 percent—all of this with a 27 percent increase in private wages." Unfortunately for this line of argument, the real minimum wage had been *falling since 1981*!

A few other simple clarifications should be stressed. The reduction in employment predicted in the standard model is not necessarily accomplished by that number (or any number) of workers being discharged; since turnover rates in minimum wage jobs are on the order of 12.5 percent *per month* (Converse et al., 1981, p. 264), attrition is a speedy method of adjustment.[2] That turnover rate also suggests that one should be careful about contrasting "those lucky enough to work at a higher wage" with "those who are unable to find work;" with high turnover rates, the two groups *may* be composed of largely the same people at different moments.

---

[1] The textbook exception of the monopsonist whose employment rises in response to a skillfully set minimum wage has little impact on the minimum wage literature outside of textbooks, in large part because the company town is not (if it ever was) the context of most minimum wage employment. I'm not sure how to summarize the attitude of the "typical" dual labor market economist on the minimum wage, but I *think* most agree it is likely to reduce employment.

Finally, unemployment due to the minimum wage as it would be measured in official statistics cannot be represented simply in the typical textbook diagram: those who would like to work at the minimum wage but are not employed may (if they actively look for work in the previous month) or may not (if they don't) be counted as unemployed by the Bureau of Labor Statistics.

A simple extension of the textbook model assumes that, while the number employed depends on the wage, the number of labor force participants (those working or looking for work) depends on the wage and the proportion of such participants who are in fact employed. Since employment also depends upon the minimum wage, the model can be solved for labor force as a function of the minimum wage alone. In general, an increase in the minimum can either increase or reduce the number of labor force participants and the unemployment (the difference between labor force and employment) which results can be either greater or less than the reduction in employment.

The simple response of the market for unskilled labor described so far leads to changes in other markets as well. Employers have an incentive to use more of other inputs, such as more skilled labor, if those inputs are substitutes for the now more expensive unskilled labor. This leads to greater employment and higher wages for such inputs when the general equilibrium response to the minimum wage is completed. Of course, the reverse occurs for inputs which are complements for unskilled workers.

Complications: Partial Coverage and Heterogeneous Workers

To interpret the empirical estimates of the effect of the minimum wage correctly, it is important to recognize two complications. First, the Fair Labor Standards Act has exempted some employers (generally small ones) from the minimum wage, with the standards of "smallness" gradually being tightened over time. In 1950, half of all private non-farm non-supervisory wage and salary workers were subject to the minimum wage; since 1977, about 5/6 have been subject. Today, the uncovered sector consists mostly of small retail trade and service employers.

Models of the impact of the minimum wage with partial coverage assume that employment in the covered sector is determined by demand, so the proportional reduction in employment in that sector is the elasticity of demand times the proportional change in the wage in the covered sector. Even if there is no response in the uncovered sector (a simple but unlikely case), the proportional change in total

---

[2] I stress this point because failure to see such discharges when the minimum wage is raised makes the untutored citizen skeptical of our wisdom on the subject. Sometimes, even the tutored are mislead: see, for example, U.S. Department of Labor (1965, p. 6), where considerable stress is placed on the tiny number of reports of layoffs and plant closings following the 1961 increase in the minimum wage.

employment (the dependent variable in the typical empirical study) will be smaller than the proportional change in employment in the covered sector.

Modeling the response of the uncovered sector to the minimum wage is complicated by the variety of ways in which workers (all of whom would prefer to work at the higher covered-sector wage) may get sorted into covered- and uncovered-sector jobs.[3] One assumption is that all those willing to work at the minimum wage apply for jobs in the covered sector, employers choose among them, and those not chosen work in the uncovered sector (if their reservation wage is less than the wage there) or do not work (if their reservation wage is greater than the wage in the uncovered sector). In this case, one can show that the minimum wage leads some of those who would otherwise have found jobs in the covered sector to be employed in the uncovered sector. This cushions the employment loss in the covered sector. Such a process, however, has distributional drawbacks (wages in the uncovered sector fall) and efficiency costs (too much labor is allocated to the uncovered sector, too little to the covered sector). Moreover, one can no longer deduce the impact of an increase in the minimum wage on the earnings of unskilled workers from a comparison of the proportional change in employment and the proportional change in the minimum wage, since the proportional increase in the average hourly wage of (both covered and uncovered) workers is less than the proportional increase in the minimum wage.

An alternative approach to modeling the allocation of workers across sectors is to assume that workers choose their sector. Those working in the uncovered sector earn W_u and have no difficulty finding work at that wage (because the uncovered sector wage is assumed flexible). Those working in the covered sector earn W_m when employed, but have some probability $P < 1$ of having such a job, equal to the ratio of covered sector employment to covered sector participants (employed plus unemployed). Ignoring risk aversion, transfer payments to the unemployed, and the distinction between new entrants and incumbents (that is, assuming unskilled jobs turn over frequently),[4] workers choose their sector based on expected earnings (wage times probability of having a job). They work in the covered sector if $PW_m > W_u$, and in the uncovered sector if $PW_m < W_u$. Equilibrium requires that $W_u = PW_m$.

Since workers care about expected earnings, and these are equal in the two sectors, the number of labor force participants is a function of W_u or, equivalently, PW_m. The total number of participants must in equilibrium equal the number employed (in both sectors) plus the number unemployed (looking for jobs in the covered sector).

As the preceding discussion implies, allowing workers to choose their sector considerably complicates the analysis. It does, however, maintain the theoretical link between the minimum wage and that unemployment which comes from queuing for minimum wage jobs. Moreover, it points out the possibility that workers will flow *from* the uncovered sector *to* the covered sector. If covered sector demand is inelastic, and workers' sectors and W_u are temporarily frozen, the minimum wage raises expected earnings in the covered sector ($PW_m > W_u$). Equilibrium can be restored only if some workers "migrate" to the covered sector, raising W_u and reducing P until $PW_m = W_u$.

---

[3] The classic two sector models of the minimum are Gramlich (1976), Mincer (1976) and Welch (1976).

[4] For treatments of more complicated cases, see Gramlich (1970), Mincer (1976), and Brown, Gilroy, and Kohen (1982).

Whatever model one chooses, the *basic* message for empirical work remains: one must account for the extent of coverage, and the elasticity of total unskilled employment with respect to the minimum will be smaller than the elasticity of demand.

The second complication for empirical work arises from the fact that minimum wage workers bear no unique identifying mark besides their wage rate. While being young, black, female, or poorly educated, or living in a low wage area will increase one's chances, any available data series on employment will combine minimum wage and higher-paid workers. Employment of higher-paid workers is not directly reduced by an increase in the minimum wage; if, for example, better-paid teenagers are substitutes for minimum wage teenagers, an increase in the minimum wage would increase better-paid teens' employment. Even if their employment is not affected at all, the elasticity of total teenage employment with respect to the minimum wage is equal to the elasticity of demand for minimum wage teens' labor times their share of teenage employment.

A more interesting and complex view emerges if labor quality or skill is modelled as continuously variable. Suppose the minimum wage rises from $3.35 to $3.85 per hour. Those previously earing $3.90 per hour, who are presumably good substitutes for those who used to earn $3.85 or less, enjoy an increase in demand for their labor. This raises their wage somewhat and suggests an upward "ripple effect" on even better-paid workers. Apart from Pettengill's (1981) work, neither proponents nor opponents of the minimum wage have given such ripples careful theoretical attention in analyzing the effect of the minimum wage on employment.

Assume a continuous distribution of workers' skills, and thus a continuous wage distribution in the absence of a minimum wage. What will happen to the wage distribution if a minimum wage is now imposed?

I suspect most readers answered that the minimum wage leads to the disappearance of the portion of the wage distribution below the new minimum, as employers discharge (or refuse to hire) those with value of marginal product below the minimum. I know I would have given that answer.

A look at an actual wage distribution gives a quite different view. Rather than being truncated at the minimum wage, there is an obvious spike at that point. As noted above, 6 percent of those employed in 1980 reported a wage equal to (actually, within five cents of) the minimum wage. If one focuses on a low-wage group, the spike is more pronounced; for example, just over a third of teenagers in retail trade and service industries earned the minimum wage in 1978 (Fritsch, 1981, p. 24), including a fifth of those working in establishments not legally subject to the minimum wage.[5]

Explaining this spike has not received much attention in the literature, and at first glance it is puzzling. Suppose workers are measured by "efficiency units," where a worker's efficiency units equal the wage that should be received in the absence of a

---

[5] In 1980, the minimum wage was $3.10, and those earning $3.05 to $3.15 were counted as receiving the minimum. In 1978, the minimum wage was $2.65, and those earning $2.65–$2.69 were counted as receiving the minimum. The count of those earning the minimum is therefore *not* boosted by counting the relatively large number of workers reporting round-dollar amounts like $3.00.

minimum wage. Now a minimum wage of $3.35 is imposed, and a spike appears in the wage distribution at that point. This means that workers who earned different wages in the absence of the minimum wage (and hence embody different numbers of efficiency units of labor) are earning the same wage. But if employers find it profitable to employ the worst of these workers at $3.35, they would prefer instead to hire the best at slightly more than $3.35. If so, to mix a metaphor, the spike should unravel.

One plausible explanation is that employers *are* paying the best of these workers more than the worst, but they are using non-wage forms of compensation. The hidden bonus (to the best) or hidden penalty (on the worst) could take several forms; changes in on-the-job training, fringes, workweek, or effort have all been suggested as possible consequences of the minimum wage, though not necessarily in the context of "the spike." Suppose, however, that employers do require greater effort from those with fewer efficiency units of labor, so that the marginal products of all those paid the minimum wage in the covered sector were equal. Why would a significant fraction of employers not bound by the law choose to pay the minimum wage, as the evidence suggests they do?[6] Whether a complicated market clearing explanation or a simpler appeal to fairness (with the minimum wage essentially defining what is fair) provides a better answer needs to be explored.

Estimated Impact on Employment and Unemployment

If policy makers are not completely certain about the size of the effect of the minimum wage on employment and unemployment, it is not for lack of trying on the part of economists—there have been over two dozen time series studies of the effect of the minimum wage on teenagers, and nearly as many other studies focused on slightly different issues, primarily cross-sectional studies of youth, low-wage areas, and low wage industries. (For a survey, see Brown, Gilroy, and Kohen, 1982.) I will focus here primarily on the time series studies of teenagers.

A typical study of this genre regresses either the employment-to-population ratio or the unemployment rate of teenagers on a minimum wage variable, an indicator of aggregate economic conditions (like the adult male unemployment rate), a time trend, and perhaps other variables such as the fraction of youth in the armed forces or the size of the teenage population relative to the entire population of labor force age.

While for convenience one refers to "the" minimum wage, quite often there are two minimum wages. When the coverage of the law is increased, and the minimum wage raised, a lower minimum wage W_m^* is sometimes introduced for the newly covered employers, which then rises toward the "basic" minimum W_m until the latter applies to all covered employers. Most studies combine the coverage and level of the minimum wage into a single minimum wage index, by weighting W_m^* and W_m by the proportion of workers to which they apply—proportions which sum to less than one because no minimum wage applies to some workers—and dividing by the average wage.

---

[6] Holzer, Katz, and Kruger (1987) are investigating whether the minimum wage creates a queue of jobseekers, as would be expected if attempts to offset the minimum wage by increasing effort and so on were only partially successful.

If one asks of the two dozen time series studies, "What is the estimated impact of a 10 percent increase in the minimum wage on teenage employment?" the answer comes back, "It falls by 1 to 3 percent." If one focuses on studies which include a time trend and control for coverage (either as a separate variable or in the combined form discussed above[7]) the estimates are 1 to 2 percent. This effect translates into an estimated increase in the teenage unemployment rate in a range from zero to 3 percentage *points*. More recent studies which include the 1970s, and studies which control for the teenage population share, produce estimates of a rise in teenage unemployment rates of .75 percentage points or less. As noted above, however, unemployment may increase most sharply in precisely the situation (inelastic covered sector demand) where average earnings of the unskilled as a group rise. So the unemployment rate deserves less focus than it gets in public discussions.

Together, the employment and unemployment effects estimated by the typical study suggest a reduction in labor force participation due to the imposition of a minimum wage. If one accepts the idea that labor force participation is an increasing function of the average rewards from participating in the labor force, it appears these rewards are reduced for teenagers by the minimum wage. However, the reduction in labor force participation does not show up as an increase in "discouraged workers," those who are not officially counted as unemployed because they have given up active job search, so the labor force withdrawal is somewhat puzzling.

If the minimum wage has negative effects on teenage employment, one would almost certainly expect to find larger (proportional) effects on employment of non-white teenagers, because a larger proportion of non-white teenagers are working at the minimum wage. Studies which address this question show very mixed results, with blacks more adversely affected than whites only about as often as the reverse, and in any case with large standard errors for the effects on blacks. If one looks to the theory for signs but the data for magnitudes, one probably should conclude that the effect of a minimum wage on black teenagers is worse than that on white teenagers, but not dramatically so.

Thus, my reading of the time series estimates is that they are not negligible—especially when 30 percent increases in the minimum wage are under discussion in Congress—but one should look elsewhere (and I am not sure where) for the primary causes of high unemployment rates in the youth labor market, particularly for black youth.

There appear to be no published studies of the consequences of the gradual repeal of the minimum wage (as rising wages elsewhere eroded a fixed nominal minimum) since 1981. An imperfect alternative, which ignores changes in age

---

[7] When coverage is entered separately, its effect tends to be imprecisely estimated—perhaps because, as noted above, there is some tendency for uncovered employers to match the minimum wage. But it's hard to see an *a priori* case for excluding it altogether, as some studies with relatively large estimates of the impact of the level of the minimum wage have done. The most detailed treatment of coverage is Al-Salam, Quester, and Welch's (1981) paper.

Table 1
Teenage Labor Market in the 1980s

Variable	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*
Minimum Wage	$3.10	$3.35	$3.35	$3.35	$3.35	$3.35	$3.35
Average Hourly Earnings (all ages)	$6.66	$7.25	$7.67	$8.20	$8.33	$8.57	$8.76
MW/AHE (%)	46.5	46.2	43.7	41.7	40.2	39.1	38.2
Unemployment Rate Aggregate (%)	7.1	7.6	9.7	9.6	7.5	7.2	7.0
Unemployment Rate Males age 20 + (%)	5.9	6.3	8.8	8.9	6.6	6.2	6.1
Employment-Population Ratio Teenagers (%)	46.6	44.6	41.5	41.5	43.7	44.4	44.6
Unemployment Rate Teenagers (%)	17.8	19.6	23.2	22.4	18.9	18.6	18.3
Employment-Population Ratio Black Teenagers (%)	23.8	21.9	19.0	18.7	21.9	24.6	25.1
Unemployment Rate Black Teenagers (%)	38.5	41.4	48.0	48.5	42.7	40.2	39.3

structure of the population, the availability of training for those not employed, and other factors held constant in more formal studies, is simply to look at what has happened to the youth labor market over this period. Some relevant figures are presented in Table 1. The first three lines show that the erosion of the minimum wage has been substantial. The next two lines show the aggregate and adult male unemployment rates, two alternative measures of the state of the overall labor market, which must be held constant in looking for effects of the minimum wage. The overall unemployment rate suggests comparing 1984 to 1981 and 1985–86 to 1980, while the adult male rate suggests treating 1984–86 and 1981 as comparable. But regardless of the comparison, this quick look at the evidence provides little support for the idea that the 20 percent reduction in the relative minimum wage had major positive effects on teenage employment; indeed, the employment-population ratio for teens is lower afterwards than it was before the minimum wage cut, while the unemployment rate comparisons are mixed and show uniformly small changes.[8] There *is* some improvement in the employment of black teens (depending on one's choice of the right comparison, it is up by as much as two points or 9 percent), though the unemployment numbers of black teens are less cheering.

The thrust of the preceding paragraphs is that the effects of the minimum wage on teenage employment and unemployment are smaller than one might have guessed (certainly smaller than I would have guessed). Less numerous studies of 20- to 24-year-olds suggest smaller proportional effects for "older youth," as one would expect from the fact that a smaller proportion of workers in this age bracket are minimum wage workers. However, two strands of evidence do suggest more noteworthy effects of the minimum wage on youth labor markets.

---

[8] Research in progress by Alison Wellington suggests that the impact of the minimum wage on youth looks *less* severe when the sample period is extended to include the 1980s.

The first strand is the somewhat fragmentary evidence that minimum wages increase the proportion of employed teenagers who are part-time workers (Gramlich, 1976).[9] The evidence is fragmentary because time series of youth employment by part- and full-time status began about ten years after the beginning of the all-teen time series used in other studies. Perhaps as a result, the size of the shift to part-time work in response to the minimum wage bounces around according to the particular estimating equation (Brown, Gilroy, and Kohen, 1983). There is complementary evidence (Mattila, 1981) that increases in the minimum wage increase school enrollment, perhaps because they increase the difference in expected earnings between dropouts and high school graduates. Since existing studies of the part-time issue used time series running from 1963 to the mid- or late 1970s, the 6 to 10 years additional data now available justify an update.

The second piece of evidence suggesting larger disemployment effects is a cross-section study by Meyer and Wise (1983).[10] In this model, the wages of young workers in the absence of the minimum wage depend on their personal characteristics and an error term. For those who would otherwise earn less than the minimum, passage of a minimum wage can do one of three things: cost them their job, leave them unaffected (if their job is not covered, or if their employer ignores the law), or result in their wage being raised up to the minimum.[11] (Recall the wage distribution's discernable spike at the minimum wage.) Loosely speaking, Meyer and Wise estimate the parameters of the wage distribution from the segment of the distribution above the minimum wage, and extrapolate that distribution back below the minimum to predict the number of young workers who would be employed at each wage below the actual minimum if there were no minimum. The difference between fitted and actual distributions below the minimum measures the extent of disemployment.[12] Their simulations suggest a 10 percent increase in the minimum wage reduces teenage employment by somewhat more than the 1 to 2 percent in the time series studies.[13] One weakness in the Meyer and Wise study for our purposes is that their estimates apply to out-of-school youth. Mattila's (1981) time series evidence suggests disemployment effects are particularly large for that group.

---

[9] The finding that higher minimum wages reduce workweeks raises an interesting theoretical issue. Employers pay higher wages to full-time than to part-time workers, suggesting that value of output per hour is higher on full-time schedules. If the observed change in workweek is a response by employers to the higher minimum wage, they should be increasing the workweek to raise the average value of output produced per hour up to the new minimum wage (Barzel, 1973). But if firms raise average output per hour by limiting on-the-job leisure ("coffee breaks"), paid hours per week would fall.

[10] I place less weight on the more traditional cross-section studies, which try to explain cross-state differences in youth employment-population ratios by a cross-state minimum wage index. With a national minimum wage (admittedly, with some variation in proportion covered due to differences in industrial structure and state minimum wage laws), most of the variation in the coverage-weighted relative minimum comes from variation in average wages across states. We might expect more youth to work in states where average wages are higher, which would produce a negative coefficient for the minimum wage variable in these studies, even if the minimum wage had no effect.

[11] The Meyer and Wise approach therefore abstracts from any effect of the minimum wage on wages below the minimum, or the possibility of upward "ripples" above the minimum.

[12] Meyer and Wise are keenly aware of the fact that their method leans heavily on the assumed functional form of the earnings function and particularly the error term. In addition to some sensitivity tests, they also estimate the wage distribution in a year with a high minimum wage and use the parameters to estimate the effect of the minimum in a low minimum wage year. The estimates line up well with what one gets using the "fuller" wage distribution from the low minimum wage year directly.

A Lower Minimum Wage for Teenagers?

If the ideal law for defenders of the minimum wage is an indexed minimum that cannot be eroded by increases in nominal wages throughout the economy, the constrained optimum for some opponents is to enact a lower minimum wage for youth. Curiously, while economists have few good things to say about minimum wage laws and support for a differential minimum wage for young workers tends to come from opponents of higher minimum wages, there seems to be little evidence of widespread enthusiasm among economists for a youth "subminimum."

The basic issues are fairly straightforward. Traditional minimum wage legislation changes the relative price of minimum wage workers, both youth and adults, relative to better-paid youth and adults. A subminimum, however, reduces the price of low wage teenagers compared to low-wage adults as well as better-paid workers of all ages. Thus, the subminimum should increase teenage employment, for the usual reasons: it makes output produced by low-wage teenagers cheaper (scale effect), and encourages employers to substitute low-wage teens for other inputs (substitution effect).

While there is little solid evidence on the point, it seems plausible that low-wage teenagers and low-wage adults are good substitutes. Perhaps they are not perfect substitutes—for example, they probably have different constraints on the hours they will work and perhaps different combinations of ability and reliability—but overall it is hard to see *a priori* how they could fail to be good substitutes. Thus, *part of* the gains of low-wage teenagers might be expected to come at the expense of low-wage adults (with the remainder coming from the scale effect and substitution against other inputs).[14] Moreover, past changes in the minimum wage law do not offer the natural experiment needed to estimate with confidence the relative size of the adult loss relative to the teen gain.

If one finds a proposal which encourages substituting low-wage teens for low-wage adults (who are at least somewhat more likely to be poor) to be ironic, one should not overlook the irony of the present law. The Fair Labor Standards Act has permitted, on a gradually more liberal basis, employers to hire full-time students working during

[13] They report that holding the nominal minimum wage at $1.60 would have increased teenage employment by 5 percent. Since the minimum wage averaged $2.21 over the four years they studied, they were in effect considering a 28 percent reduction. However, this brings the minimum wage well below the historical range compared to average hourly earnings, and the elasticity of employment with respect to the minimum falls in absolute value as one considers successively lower values of the minimum.

[14] Whether a youth subminimum is efficient (given a minimum wage in the first place) is unclear. While it reduces the underutilization of unskilled youth, it intensifies the underutilization of unskilled adults.

the summer (or part-time during the school year) at 85 percent of the minimum wage. Here, the substitution which seems to be encouraged is to hire (advantaged?) youth still in school instead of (less advantaged?) dropouts. Mercifully, even in 1979 (a high-use year) students hired under this subminimum accounted for less than 5 percent of student employment, and there is no evidence of substitution effects exceeding scale effects for non-student youth (Freeman, Gray, and Ichniowski, 1981).

There is a further irony in the debate over a differential minimum wage for youth. Earlier, I noted that an absence of evidence of widespread discharges following minimum wage increases had led some supporters of the minimum to doubt that it was causing any significant loss of employment. The same logical error seems to have crept into the thinking of those critics of the minimum wage who support a youth differential. To respond to charges that a youth differential will lead employers to substitute now-cheaper youth for adults, proposed youth differential bills invariably include a provision making it illegal for employers to discharge adults in order to hire teenagers. The existing subminimum for students has a similar provision, and it is a "dead letter" (Freeman, Gray, and Ichniowski, 1981, p. 309).

Effect of the Minimum Wage on the Income Distribution

If the minimum wage is increased by 10 percent, roughly a quarter of teenagers (closer to half if one counts those below the minimum) get a 10 percent increase in their wage, while (according to the time series estimates) 1 to 2 percent are no longer employed. Unless wages fall precipitously in the uncovered sector, or hours reductions are large, teens as a group seem to gain.[15] If one can get past these qualifications, and assumes that, proportionally, things work about the same for low-wage adults as for teenagers, one might see an iffy case for an increase in the minimum wage to reduce poverty.

One of the more surprising findings in the minimum wage literature is that, even under the most favorable assumption that the minimum wage had *no* disemployment effect, its effect on poverty or the income distribution is not very large. The reason is the surprisingly weak relationship between being a worker whose hourly wage is low and being a member of a family whose annual income is low.[16] Many poor families have no earners at all, or earners whose wage is above the minimum but cannot work enough hours per year. While there are a number of ways of expressing this correlation, Johnson and Browning's (1983) simulation is perhaps the most dramatic. They find that in 1976, when the minimum wage was $2.30, earnings of workers making less than *$2.80* per hour accounted for only 11 percent of the after-tax, after transfer income of the poorest fifth of all households. Consequently, even a 22 percent

---

[15] Meyer and Wise show that, assuming work weeks and wages below the minimum are unaffected, their (high-side) effects on employment leave teenagers' earnings unaffected by the minimum wage.

[16] Some evidence on the weakness of this correlation was first presented by Stigler (1946). More recent studies, with more broad-based samples, include Kelly (1976), Gramlich (1981), Bell (1981), and Kniesner (1981). Burkhauser and Finegan (1987) show that the correlation was once considerably stronger.

increase in the minimum wage with no loss of employment raises the income of these households by less than one percent.

To be fair, the finding that low-wage workers are often not members of low-income households ought to temper the claims of opponents of the minimum wage that it is a particularly bad idea for the poor. But if one is to overcome an economist's natural reluctance to mess with markets by pointing to distributional gains, the limited nature of these gains under even favorable assumptions about employment effects leaves one grasping (and perhaps gasping) for other tools emphasizing regular full-time employment and/or direct transfers.

Conclusions and Regrets

And what, after these pages of findings and qualifications, do I think we've learned? The effects of the minimum wage on employment are smaller than I would have supposed. Part of this small effect is due to the fallacy of the inflated denominator: with incomplete coverage and the fact that most workers in nearly any demographic group earn more than the minimum wage, the effect on any such group's employment will be smaller (in proportional terms) than it would be if impacts on those directly affected could be isolated. If one comes to the issue willing to suffer some efficiency loss for a real distributional gain—at least until someone isolates a lump-sum grant—than the real surprise is not the disemployment effects but the very tenuous connection of the minimum wage to poverty. It is hard for me to see evidence that minimum wage increases have benefits which would overcome an economist's aversion to interfering with reasonably competitive markets. But the case against the minimum wage seems to me to rest more upon that aversion than on the demonstrated severity of any harm done to those directly affected.

A short paper like this one must skip some issues. In a longer paper, the short literatures on whether low-wage employers reduce the training they provide in response to the minimum, on how effectively the minimum wage is enforced, on whether the minimum wage in fact has an appreciable upward ripple effect on other wages, and on who supports and opposes minimum-wage legislation—and why—would deserve more attention. But answers to those questions do not seem likely to change the basic conclusion that the minimum wage is overrated: by its critics as well as its supporters.

References

Al-Salam, Nabeel, Aline Quester, and Finis Welch, "Some Determinants of the Level and Composition of Teenage Employment." In Rottenberg, Simon, ed., *The Economics of Legal Minimum Wages*. Washington, DC: American Enterprise Institute, 1981, pp. 124–154.

Barzel, Yoram, "The Determinants of Daily Hours and Wages," *Quarterly Journal of Economics*, May 1973, *87*: 220–238.

Bell, Carolyn Shaw, "Minimum Wages and Personal Income." In Rottenberg, Simon, ed., *The Economics of Legal Minimum Wages*. Washington, DC: American Enterprise Institute, 1981, pp. 429–458.

Brown, Charles, Curtis Gilroy, and Andrew Kohen, "The Effect of the Minimum Wage on Employment and Unemployment," *Journal of Economic Literature,* June 1982, *20,* 487–528.

Brown, Charles, Curtis Gilroy, and Andrew Kohen, "Time-Series Evidence of the Effect of the Minimum Wage on Youth Employment and Unemployment," *Journal of Human Resources,* Winter 1983, *18,* 3–31.

Burkhauser, Richard, and T. Aldrich Finegan, "The Minimum Wage and the Poor: The End of a Relationship," unpublished paper, 1987.

Converse, Muriel, et al., "The Minimum Wage: An Employer Survey." In Minimum Wage Study Commission *Report,* Vol. 6. Washington DC: U.S. G.P.O., 1981, pp. 241–341.

Eccles, Mary, and Richard Freeman, "What! Another Minimum Wage Study?" *American Economic Review,* May 1982, *72,* 226–232.

Freeman, Richard, Wayne Gray, and Casey Ichniowski, "Low Cost Student Labor: The Use and Effects of the Subminimum Wage Provisions for Full-Time Students." In Minimum Wage Study Commission *Report,* Vol. 5. Washington DC: U.S. G.P.O., 1981, pp. 305–335.

Fritsch, Conrad, "Exemptions From the Fair Labor Standards Act, Retail Trade and Services." In Minimum Wage Study Commission *Report,* Vol. 4. Washington DC: U.S. G.P.O., 1981, pp. 1–96.

Gilroy, Curtis, "A Demographic Profile of Minimum Wage Workers." In Minimum Wage Study Commission *Report,* Vol. 2. Washington, DC: U.S. G.P.O., 1981, pp. 153–213.

Gramlich, Edward, "The Impact of Minimum Wages on Other Wages, Employment, and Family Incomes," *Brookings Papers on Economic Activity,* 1976 : 2, 409–451.

Holzer, Harry, Lawrence Katz, and Alan Krueger, "Job Queues and Wages: New Evidence on the Minimum Wage and the Inter-Industry Wage Structure," unpublished paper, 1987.

Johnson, William, and Edgar Browning, "The Distributional and Efficiency Effects of Increasing the Minimum Wage: A Simulation," *American Economic Review,* March 1983, *73,* 204–211.

Kelly, Terrence, "Two Policy Questions Regarding the Minimum Wage," Working Paper No. 3608-05, The Urban Institute, 1976.

Knieser, Thomas, "Low-Wage Workers: Who Are They?" In Rottenberg, Simon, ed., *The Economics of Legal Minimum Wages.* Washington D.C.: American Enterprise Institute, 1981, pp. 459–481.

Mattila, J. Peter, "The Impact of Minimum Wages on Teenage Schooling and on the Part Time/Full Time Employment of Youths." In Rottenberg, Simon, ed., *The Economics of Legal Minimum Wages.* Washington DC: American Enterprise Institute, 1981, pp. 61–87.

Meyer, Robert, and David Wise, "The Effects of the Minimum Wage on the Employment and Earnings of Youth," *Journal of Labor Economics,* January 1983, *1,* 66–100.

Mincer, Jacob, "Unemployment Effects of Minimum Wages," *Journal of Political Economy,* August 1976, *84,* S87–S105.

Pettengill, John S., "The Long-Run Impact of a Minimum Wage on Employment and the Wage Structure." In Minimum Wage Study Commission *Report,* Vol. 6. Washington, DC: U.S. G.P.O., 1981, pp. 63–104.

Smith, Ralph E., and Bruce Vavrichek, "The Minimum Wage: Its Relation to Incomes and Poverty," *Monthly Labor Review,* June 1987. **110,** 24–30.

Stigler, George, "The Economics of Minimum Wage Legislation," *American Economic Review,* June 1946, *36,* 358–365.

Subcommittee on Labor Standards, Committee on Education and Labor, U.S. House of Representatives. [Hearings on the] *Fair Labor Standards Amendments of 1977.* Washington, DC: U.S. G.P.O., 1977.

Subcommittee on Labor Standards, Committee on Education and Labor, U.S. House of Representatives. *Hearings on H.R. 1834, The Minimum Wage Restoration Act of 1987.* Vol. 1, Washington DC: U.S. G.P.O., 1987.

Welch, Finis, "Minimum Wage Legislation in the United States," In Ashenfelter, Orley, and James Blum, eds., *Evaluating the Labor Market Effects of Social Programs.* Princeton, NJ: Princeton University Press, 1976, pp. 1–38.

U.S. Department of Labor, "Report Submitted to the Congress in Accordance with the Requirements of Section 4(d) of the Fair Labor Standards Act," January 1965.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 147–174

Variable Trends in Economic Time Series

James H. Stock and Mark W. Watson

The two most striking historical features of aggregate output are its sustained long run growth and its recurrent fluctuations around this growth path. Real per capita GNP, consumption and investment in the United States during the postwar era are plotted in Figure 1. Both growth and deviations from the growth trend—often referred to as "business cycles"—are apparent in each series. Over horizons of a few years, these shorter cyclical swings can be pronounced; for example, the 1953, 1957 and 1974 recessions are evident as substantial temporary declines in aggregate activity. These cyclical fluctuations are, however, dwarfed in magnitude by the secular expansion of output. But just as there are cyclical swings in output, so too are there variations in the growth trend: growth in GNP in the 1960s was much stronger than it was in the 1950s. Thus, changes in long run patterns of growth are an important feature of postwar aggregate economic activity.

In this article we discuss the implications of changing trends in macroeconomic data from two perspectives. The first perspective is that of a macroeconomist reassessing the conventional dichotomy between growth and stabilization policies. As an empirical matter, does this dichotomy make sense for the postwar United States? What is the relative "importance" of changes in the trend and cyclical swings in explaining the quarterly movements in economic aggregates? We next adopt the perspective of an econometrician interpreting empirical evidence based on data that contain variable trends. The presence of variable trends in time series data can lead one to draw mistaken inferences using conventional econometric techniques. How can these techniques—or our interpretation of them—be modified to avoid these mistakes?

■ *James H. Stock is Assistant Professor of Public Policy, Kennedy School of Government, Harvard University, Cambridge, MA. Mark W. Watson is Associate Professor of Economics, Department of Economics, Northwestern University, Evanston, IL. This article was written while Stock was a National Fellow at the Hoover Institution, Stanford University, Stanford, CA.*

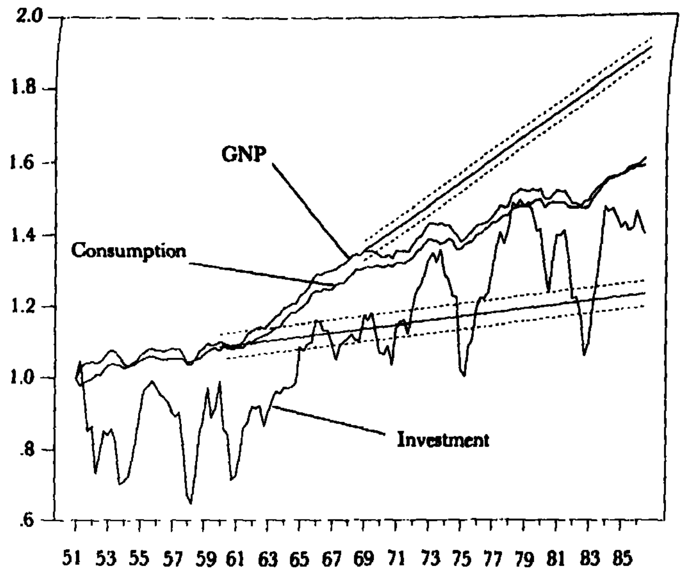


Fig. 1. Postwar real per capita U.S. GNP, total consumption, and gross private domestic investment (in logarithms)

All three series have been arbitrarily set to 1 in 1951:I. The straight solid lines represent two long-run forecasts of GNP, the lower using data from 1951:I–1959:IV and the upper using data from 1960:I–1969:IV. The forecasts were made by extrapolating GNP growth over these periods using a linear deterministic time trend. The dotted lines represent bands of ± two standard deviations of quarterly GNP growth around the long-run forecasts. Were GNP a stationary series about a linear time trend, these bands would provide an approximate long-run 95% confidence interval for the respective GNP forecasts.

The macroeconomist's perspective is adopted in the first major section of the article. Discussions of macroeconomics often treat the concepts of "trends" and "cycles" in output separately. On the one hand, theories of growth examine the forces capable of changing long run trends, while on the other hand theories of the business cycle attempt to explain shorter run fluctuations and to determine when macroeconomic policy might stabilize or exacerbate the swings between expansion and recession. At one level this dichotomy seems natural, with different theories providing insights into macroeconomic movements over different horizons. But on another level this distinction is artificial: theories explaining only growth or only cycles cannot provide adequate macroeconomic insights if there are important interactions between the two.

A closer look at Figure 1 suggests that changes in growth trends are associated with some of the shorter, "cyclical" swings in the series. For example, a key turning point between the high-growth 1960s and the low-growth 1970s and 1980s seems to have been in the early 1970s, which was also followed by a major recession that saw a particularly sharp transitory drop in investment. When one formally defines trends using the "stochastic trends" concept discussed below, various statistical measures (introduced in the next section) indicate that a substantial fraction of the quarterly variation in real GNP is associated with movements in long run trends rather than being purely transitory fluctuations. This agrees with the casual inference drawn from Figure 1 that shifts in trends are an important part of changes in GNP. Moreover, examinations of the co-movements of several macroeconomic variables indicate that unforecasted shifts in long run economic prospects are associated with short run fluctuations that are similar to common conceptions of the business cycle. These findings suggest two conclusions: first, that a key step in understanding the co-movements in aggregate economic variables is learning about the link between changes in long run economic trends and cyclical swings; and second, that a sharp dichotomy between growth and stabilization policies misses an important connection between the two policy goals.

The econometrician's perspective on variable trends is taken in the second major section of this article. Variable trends provide numerous econometric pitfalls and raise difficult methodological issues. Time series analysts have long recognized that regression analysis can be highly misleading when applied to series with variable trends. In some cases (as in Figure 1), the result can be dramatic errors in forecasting. In other cases, an improper treatment of variable trends can result in false conclusions about how the economy works. To illustrate this point, we construct a simple artificial economy and assign two hypothetical inhabitants the task of discovering its true structure. As in many model economies studied in the modern literature on macroeconomic theory, these inhabitants happen to be econometricians; but here their training often fails them simply because they mishandle the variable trends in their data. Finally, we draw on recent developments in econometric theory to provide some simple rules-of-thumb that a consumer or producer of econometric models can use in an attempt to avoid these pitfalls.

What is a "Variable Trend?"

The basic premise of econometric analysis is that, when viewed together, individual cases and experiences can provide insights into a deeper unifying structure. When analyzing cross sectional data, the individual experiences might be those of different workers or firms; but the only individual experiences we have with the operation of the U.S. economy are historical. For macroeconometric analysis to be of any value, then, it must be that the historical experiences comprising an economic time series are on the one hand sufficiently different from each other that more experiences provide additional information, but on the other hand sufficiently similar that combining individual experiences can elucidate the underlying economic structure. These two requirements are the essence of the technical assumption that a time series is ergodic and stationary.[1]

The first requirement—that historical experiences be sufficiently unrelated—is unlikely to hold for GNP, consumption and investment in Figure 1. The key reason is that the current trend levels of these variables arguably depend on the entire history of these series. For example, were the experiences of the 1950s essentially unrelated to those of the 1980s, except for a perfectly predictable trend, an econometric forecaster in 1960 might predict 1987 GNP by extrapolating trend growth during the 1950s out to 1987; an econometrician in 1970 might do the same, only using data from the 1960s. With unchanging trends, these forecasts should be similar, and indeed should be moderately accurate. But, as indicated in Figure 1, these forecasts are close neither to each other nor to the actual value of 1987 GNP. Just as important, the confidence bands around the trend extrapolations in Figure 1—designed to capture the uncertainty associated with these forecasts—clearly fail to provide a reliable estimate of the range of future GNP. Thus historical experiences—namely, unforecasted increases in the level of GNP in some past year—appear to have had a dramatic and persistent influence on production today.

---

[1] Formally, a time series random variable is said to be stationary if its distribution does not depend on time. As is the convention, we further assume that *stationary* variables have finite variances and autocovariances. See Andrew Harvey (1981, p. 22) or Clive Granger and Paul Newbold (1977, p. 4) for more details.

On a casual level, the notion that GNP is composed of a variable trend plus some additional cyclical movements seems quite satisfactory. Time series variables can, without further restrictions, be thought of as composed of a part with a variable trend, plus a part that is not the trend. But a moment's reflection indicates that this decomposition lacks content; without a more precise definition, one economist's "trend" can be another's "cycle." The ambiguity surrounding the definition of a variable trend applies equally to the definition of a "business cycle." Indeed, one can imagine extreme views that there is no business cycle, in the sense that all economic fluctuations are merely movements in variable trends—or the reverse view that what appear to be variable macroeconomic trends are but very long cyclical swings. But these extreme views miss valuable notions traditionally associated with the business cycle: that the economy experiences protracted but nonetheless temporary periods of unusually high (or low) output, employment, and inflation across many or all sectors of the economy; that these protracted experiences have important implications for the well-being of people and institutions; and that these changes in well-being often have important political consequences. In referring to business cycles, then, we mean these protracted yet temporary swings in aggregate output.

Then what is a trend? Perhaps the theory of economic growth can provide an operational definition. For example, in a one-sector neoclassical growth model, trends can arise because of technical progress and an increasing labor force and capital stock. If the production technology has constant returns to scale, this suggests that—at least in theory—the trend in output per capita could be represented as a function of the capital-labor ratio, the labor force participation rate, and the stock of "technology." Unfortunately, in practice things are not so simple; for example, the capital-labor ratio and the labor force participation rate have cyclical as well as trend components, and the stock of technology cannot be measured directly. Thus, at least at this level of analysis, this approach based on growth theory has little empirical relevance.

It is therefore common to take a different approach to the definition of a variable trend. The specific notion we adopt is a direct extension of a *deterministic linear time trend* (used to compute the forecasts in Figure 1), which increases by some fixed amount (say, 1 percent) every quarter. In contrast, we model a variable trend as increasing in each quarter by some fixed amount (say, 1 percent) on average; however, in any given quarter the change in the trend will deviate from its average by some unforecastable random amount. Because it has this unpredictable random component, henceforth we refer to this formulation of a variable trend as a *stochastic trend*. The reader familiar with the random walk theory of stock prices will recognize that this notion of a trend corresponds to a random walk with drift. From a forecasting perspective a key feature of a random walk—in contrast to a stationary time series variable—is that, because of its random growth, forecasts of its level will entail uncertainty that increases as the forecast horizon recedes. With this formulation of a stochastic trend, a (random) change in the trend in one quarter provides a new base from which growth will occur in the next.[2]

While this definition of a stochastic trend might at first seem restrictive, a recent theorem by Stephen Beveridge and Charles Nelson (1981) suggests that it might be broadly applicable to U.S. data. Since George Box and Gwilym Jenkins (1970) proposed their influential autoregressive-integrated-moving average (ARIMA) models for forecasting time series variables, econometricians have generally recognized that many macroeconomic series (when considered one at a time) appear to be *integrated*, that is, their first differences are stationary. ARIMA models are sophisticated yet simple tools for forecasting a single series using only its history. Box and Jenkins referred to integrated variables with autoregressive (AR) or moving average (MA) components as ARIMA$(p, 1, q)$ processes, where p and q, respectively, denote the orders of the autoregressive and moving average terms and where "1" indicates that the variable is integrated of order one, which means that its first difference (quarterly change) is stationary. ARIMA models are extensions of conventional regression models to time series variables; the AR terms forecast the current variable using p of its lags, and the MA terms forecast using q lags of the error in the process.[3]

---

[2] A time series variable x_t is a *random walk with drift* if x_t evolves according to

$$x_t = \mu + x_{t-1} + e_t \text{ or } x_t - x_{t-1} = \mu + e_t$$

where e_t has mean zero and variance σ_e^2, and where e_t is serially uncorrelated, i.e. e_t is uncorrelated with e_s for $s \neq t$, so e_t cannot be forecasted using past values of x_t. The "drift" in the random walk is μ, the average predictable increase in x_t in each period. (For example, from 1951:I to 1986:III, real GNP increased by 3.0 percent annually; on a per capita basis, the average annual increase was 1.7 percent). With this definition, a variable y_t that contains a stochastic trend can be written as $y_t = y_t^p + y_t^s$, where y_t^p is a random walk (possibly with drift) and y_t^s is a stationary time series variable; the superscripts "p" and "s" refer to the "permanent" (or trend) and "stationary" (or transitory) components of y_t. In contrast, a variable that contains a deterministic time trend can be written as $y_t = gt + y_t^s$, where y_t^s is stationary and g is the constant quarterly growth of the deterministic trend.

[3] *Some Useful Definitions*

(i) The variable x_t is *integrated of order one* (or simply *integrated*) if it is nonstationary and can be written,

$$x_t = \mu + x_{t-1} + u_t \text{ or } x_t - x_{t-1} = \mu + u_t$$

where u_t has mean zero and variance σ_u^2, and where u_t is stationary. It is convenient to let Δx_t represent $x_t - x_{t-1}$. A variable is said to be *integrated of order d* if it must be differenced d times to be stationary; for example, if Δx_t is not stationary but $\Delta x_t - \Delta x_{t-1}$ is, then x_t is integrated of order 2. The distinction between an integrated process and a random walk is that, if x_t is integrated, u_t is stationary but might be correlated with lagged u_t; if x_t is a random walk, u_t is serially uncorrelated.

Among time series analysts, a major attraction of ARIMA($p, 1, q$) models is their ability to forecast many macroeconomic variables with an accuracy that is impressive among univariate forecasting techinques. Here, however, the importance of ARIMA models is that Beveridge and Nelson prove that every variable having an ARIMA($p, 1, q$) representation contains a random walk stochastic trend. Since ARIMA($p, 1, q$) models seem to characterize many macroeconomic variables, it follows that the growth in these variables can be described by stochastic trends. Beveridge and Nelson's (1981) trend/cycle decomposition is presented in the Appendix for readers familiar with the mathematical particulars of ARIMA models.

A possible objection to the discussion so far is its emphasis on variables that are integrated of order one, so that the quarterly percentage growth in GNP is modeled as being stationary. Why not treat U.S. data as integrated of order two, so that the second difference of GNP is stationary but the quarterly growth rate itself contains a stochastic trend? Or why not model GNP as stationary but having coefficients that imply almost the same "stochastic trend" behavior as an integrated process? Our answer is that the preponderance of evidence currently suggests that the integrated model provides the best approximation of U.S. GNP. This is not to say that U.S. GNP is an integrated process, for this can never be learned with certainty by examining a finite time series; nor is it to say that future research using new techniques or more data could not change this assessment. But, given currently available statistical techniques, modeling GNP as an integrated process seems to provide a good approximation to its long run properties.[4]

(ii) Following Box and Jenkins, an *ARIMA*($p, 1, q$) model specifies x_t as being integrated of order one and as having a representation of the form,

$$\Delta x_t = c + a_1 \Delta x_{t-1} + \cdots + a_p \Delta x_{t-p} + e_t + b_1 e_{t-1} + \cdots + b_{t-q} e_{t-q}$$

where $a_1, \ldots, a_p, b_1, \ldots, b_q$, and c are constant parameters and where e_t is serially uncorrelated.

(iii) A variable that is integrated is said to have a *unit root in its autoregressive representation*. The term "unit root" refers to the unit coefficient on x_{t-1} in the formula defining an integrated process. The statements "x_t has a unit root" and "x_t is integrated of order one" are equivalent. Box and Jenkins use the term "nonstationary" to refer to an integrated process. This terminology, while conventional, is unfortunate: while all integrated processes are nonstationary, not all nonstationary processes are integrated.

(iv) *An integrated process has a variance that tends to infinity.* This is most easily demonstrated for a random walk. Let $x_t = x_{t-1} + e_t$ where e_t is serially uncorrelated and $\mathrm{var}(e_t) = \sigma_e^2$, and let $x_0 = 0$. Then $x_t = \sum_{s=1}^{t} e_s$, so that $\mathrm{var}(x_t) = t\sigma_e^2$, which tends to infinity with t.

[4] Guy Orcutt (1948) was among the first to suggest that GNP is an integrated process. Using annual U.S. data from 1919 to 1932, he found that many aggregate time series were well described by an ARIMA(1, 1, 0) model with an autoregressive coefficient of .3. This argument for the existence of stochastic trends in macroeconomic data is only partly convincing, since Orcutt's and Box and Jenkins' techniques for determining whether the process is integrated require the practitioner to make qualitative judgements. To overcome this drawback, Wayne Fuller (1976) and David Dickey and Fuller (1979) proposed several statistical tests for whether a variable is integrated, against the alternative that it is not (i.e. it is stationary). Nelson and Charles Plosser (1982) provided firm support for what Orcutt and the Box-Jenkins practitioners had suspected: upon applying the formal Dickey-Fuller tests to fourteen annual macroeconomic variables using 60 and 100 years of data, they could not reject the hypothesis that there is a stochastic trend in real and nominal output measures, wages, prices, monetary variables, and asset prices. In their analysis, only the unemployment rate failed to contain a stochastic trend over the twentieth century. Nelson and Plosser also emphasized that the presence of stochastic trends calls into question the validity of the traditional trend/cycle dichotomy used to described aggregate time series.

The general success of ARIMA modeling therefore provides a technical motivation for the stochastic trend formulation. But the real motivation is provided by Figure 1, where the long-run forecasts based on a deterministic linear trend are simple, intuitively appealing, and wrong. With a deterministic linear trend, the uncertainty associated with a long run forecast is limited to the variation in the stationary deviations from that trend. In contrast, with a stochastic trend, the greater the forecasting horizon, the greater is the uncertainty associated with that forecast.

Relations Between Trends and Cycles in Macroeconomic Variables

Implicit in many models of the business cycle is the notion that macroeconomic fluctuations are, for the most part, caused by temporary rigidities or misperceptions. These models abstract from economic growth, operating on the implicit assumption that the growth process has little impact on the business cycle. In contrast, an alternative view (exposited clearly by Edward Prescott, 1987) explains economic fluctuations entirely as a reaction to changes in the long run growth prospects of the economy, with "business cycles" arising simply as adjustments to new long-run growth paths.

But theoretical debates linking changes in trends to cyclical fluctuations are of little practical interest unless empirical evidence suggests that there is such a link. In this section, we consider two related questions concerning the quantitative importance of changes in long run prospects for short run economic fluctuations. First, to what extent are quarterly movements in postwar U.S. real per capita GNP associated with variations in its trend? Second, are variations in macroeconomic trends linked to cyclical movements, or are the trend and cyclical variations largely unrelated?

Measuring the Trend in GNP

The first question seems simple enough, especially if we hold ourselves to the "random walk" concept of stochastic trends. It has, however, generated a heated debate over the importance of the trend component in GNP. Besides being of interest in its own right, this debate highlights central conceptual difficulties that arise in defining trends.

The starting point for this analysis is to suppose that real GNP consists of two parts: a stochastic trend, plus a part that is transitory, or more precisely, stationary. How important, then, are these two components? Answering this question requires first determining whether the unforecastable changes (technically, the innovations) in the two components are correlated, and second, developing an econometric framework (that is, a model) for interpreting the information contained in historical GNP data.

One would hope that an examination of historical data might shed some light on the correlation between the permanent and stationary innovations. Unfortunately this hope is vain, since this correlation cannot be estimated directly from a single time series; that is, this correlation is not identified in the usual econometric sense. This lack of identification should not be surprising. Broadly speaking, determining the correlation between the trend and stationary innovations in GNP is much like deciding whether the 1975 downturn was a result of a permanent shift in the trend, a transitory fluctuation, or some combination of the two—using only the plot of GNP in Figure 1. Without embarking on a review of the literature, we follow previous authors and consider two extremes: that these innovations are either uncorrelated or perfectly correlated. In the context of the trend-stationary decomposition, assuming the innovations to be uncorrelated implies that the changes in the trend and the transitory fluctuations are unrelated, except of course that they both affect GNP. In contrast, assuming the innovations to be perfectly correlated implies that they arise from the same source.

Although the decision concerning the correlation between the trend and stationary innovations is conceptually distinct from the decision about which model to estimate, in practice the first decision suggests a class of models to choose from. In the introduction, we emphasized that Beveridge and Nelson's theorem shows that all ARIMA($p, 1, q$) models imply the presence of a stochastic trend. In fact, their theorem does more than this: it provides an explicit formula for computing a stochastic trend implied by an ARIMA model. An ARIMA model reduces all unforeseen economic events into a single innovation, and the Beveridge-Nelson trend and stationary components are both based on this innovation.

Although they sound complicated, the ARIMA framework and Beveridge-Nelson decomposition are in fact simple. As a concrete example, consider an ARIMA(0, 1, 1) model fit to GNP. Let y_t denote the logarithm of real U.S. GNP, let $\Delta y_t = y_t - y_{t-1}$ denote quarterly real GNP growth, let e_t denote an unobserved error term, and let *SE* denote the standard error of the estimate of Δy_t, that is, the standard deviation of the in-sample one-step-ahead forecast error. Because y_t is assumed to be an integrated process, the model is specified using the stationary growth rates, Δy_t. Using data from 1947:II to 1985:IV, we estimated the ARIMA(0, 1, 1) model,[5]

$$(1) \qquad \Delta y_t = .008 + e_t + .3e_{t-1}, SE = .0106.$$

In this ARIMA(0, 1, 1) specification, GNP growth is expressed as a weighted moving average of the errors, e_t. According to (1), given e_{t-1} but not e_t, GNP growth would thus be forecasted by $.008 + .3e_{t-1}$. In addition, the Beveridge-Nelson decomposition is particularly easy to apply in this case: using (1) and (A.2) in the Appendix, the permanent, or trend, component in log output (call this y_t^p) is $y_t^p = 1.3\Sigma_{s=1}^{t} e_s$. The transitory, or stationary, component in log output (call this y_t^s), also given by (1) and (A.2), is $y_t^s = -.3e_t$. Evidently the innovations in the Beveridge-Nelson permanent and stationary components are both e_t, so that the permanent and stationary innova-

---

[5] Estimation was performed using the econometrics software package RATS. Let RGNP be the name used for the series, real GNP, in a RATS session. The ARIMA(0, 1, 1) model was estimated using the RATS commands:

```
set dlrgnp 47:2 85:4 = log(rgnp(t)) - log(rgnp(t - 1))
boxjenk(ar = 0, ma = 1, constant) dlrgnp 47:2 85:4
```

tions are perfectly correlated. Although this example is an ARIMA(0, 1, 1) model, their decomposition applies more generally to all ARIMA(p, 1, q) models. In the "perfect-correlation" case then, the choice of model reduces to a choice of p and q.

The choice of a specific model to estimate in the "zero-correlation" case is conceptually similar. In this class of models, GNP is explicitly represented as the sum of its permanent component, modeled as a random walk with drift, and its transitory component, modeled as a stationary ARMA process (i.e. as an ARIMA(p, 0, q) process). Since neither component is observed directly (only their sum, GNP, is), this model is called an unobserved components ARIMA (UC-ARIMA) model. As a specific example, Watson fit a simple UC-ARIMA model to log real GNP, where the stationary component was assumed to be a second order autoregression (i.e. to be ARMA(2, 0)). Let e_t^s and e_t^p respectively be the innovations in the stationary and permanent components. His model, estimated using data from 1949:I–1984:IV, is:[6]

$$y_t = y_t^p + y_t^s, \tag{2}$$

$$y_t^p = .008 + y_{t-1}^p + e_t^p, \text{std. dev.}(e_t^p) = .0057$$

$$y_t^s = 1.5y_{t-1}^s - .6y_{t-2}^s + e_t^s, \text{std. dev.}(e_t^s) = .0076$$

$$\text{cov}(e_t^s, e_t^p) = 0, \ SE = .0099.$$

The innovations e_t^s and e_t^p are uncorrelated by assumption. Summarizing, in (2) the permanent component is written as a random walk with a drift of .008, and the stationary term is predicted using two of its lags.

Given a choice of model, the question becomes how to measure the extent to which quarterly movements in GNP are associated with variations in its trend. Since there is no one answer, various measures addressing different aspects of this question are presented in Table 1. The results pertain to different models based on the two assumptions about the correlation between the trend and stationary innovations. The results in the first row are for the ARIMA(0, 1, 1) model reported in (1). The next two results are for other low-order ARIMA models of GNP of the type estimated by John Campbell and Gregory Mankiw (1987a) and Watson (1986), while the following two are for higher order autoregressive models of the type studied by John Cochrane (1986). The final model is a UC-ARIMA model estimated by Peter Clark (1987a) with a stochastic trend plus a stationary AR(2) component, in which the innovations in the trend and stationary components are uncorrelated by assumption.[7]

[6] Estimation of the parameters of UC-ARIMA models is more involved than for ARIMA models. Harvey (1985) provides a complete discussion of specification and estimation of UC-ARIMA models.

[7] Model 3 is taken from Campbell and Mankiw (1987a) and Model 6 is taken from Clark (1987a). Cochrane used annual data for his empirical work; the high-order autoregressive models reported here capture his notion of including many lags in an annual specification. Models 1 and 2 were also reestimated using the full data set. All models produce similar short run forecasts and fit the data well using standard time series diagnostic measures.

Table 1
Measures of the importance of the trend in real log GNP, estimated using data from 1947:I to 1985:IV

Univariate Statistical Model	*Long-run increase in GNP predicted from a 1% unforeseen increase in GNP in one quarter*	*Variance ratios* $R^2_{\text{corr}=1}$	$R^2_{\text{corr}=\text{min}}$
1. ARIMA(0, 1, 1)	1.3	.93	.73
2. ARIMA(1, 1, 0)	1.6	.86	.60
3. ARIMA(2, 1, 2)	1.5	.80	.70
4. ARIMA(12, 1, 0)	1.3	.75	.57
5. ARIMA(24, 1, 0)	.9	.79	.64
6. UC-ARIMA	.6	.84	.34

The univariate estimates are based on analysis of GNP. As discussed in the text, the UC-ARIMA model was estimated under the assumption that the trend and stationary innovations are uncorrelated. The R^2 statistics measure the fraction of the variance in the quarterly change in real GNP attributable to changes in its stochastic trend. The minimal correlations used to compute the $R^2_{\text{corr}-\text{min}}$ statistic for the univariate models are (in order): .85, .9, .95, .9, .8, 0.

Source: authors' calculations, drawing on Campbell and Mankiw (1987a); Clark (1987a); Cochrane; and Gagnon.

The first measure of the importance of the trend component in GNP addresses the following question (ennuciated most clearly by Campbell and Mankiw, 1987a): supposing that in some quarter GNP were to increase by 1 percent above its forecasted amount, how would that change one's forecast of the long run level of GNP? If GNP were a stochastic trend with no stationary component, then the answer would be 1 percent: the best forecast of a random walk arbitrarily far in the future is its current value, and if that value were to change by 1 percent, so would the long run forecast. In contrast, if GNP were stationary around a purely deterministic time trend, then the answer would be zero percent: any unpredictable change would have only a transitory effect on forecasts of future GNP.

It turns out that the choice of model by the various researchers makes a big difference in estimating the long run effect of an innovation in the trend; the estimates vary by a factor of two. Although this is unsatisfying, it is useful to understand why these estimates vary. Consider first the estimates based on the assumption of a perfect correlation between the trend and stationary innovations. Using the results in the first row of Table 1, suppose that GNP were to grow by an unforecasted 1 percent in some quarter. This growth will, on average, arise partly from an innovation in the trend and partly from innovations in the stationary component. Since the stationary innovation is perfectly correlated with the trend innovation, the change in the stationary component will either augment or partially offset the permanent innovation, respectively depending on whether the trend innovation induces an increase or a decrease in the stationary component. Using the ARIMA(0, 1, 1) model (1), since $y_t^p = 1.3\Sigma_{s=1}^t e_s$ and $y_t^s = -.3e_t$, an increase in the trend is associated with a decrease in the stationary component using the Beveridge-Nelson decomposition: for a 1.3

percent increase in the trend, the stationary component would initially drop by .3 percent, leaving a net unforecasted increase in GNP of 1 percent. Eventually the effect of this innovation on the stationary component will vanish, leaving only the 1.3 percent permanent increase.

In contrast, when an unforecasted increase in GNP consists of the two uncorrelated innovations, less of an unforecasted quarterly change will be attributed to shifts in the trend. A 1 percent increase in GNP might have come from a trend or a stationary innovation; since these two innovations on average neither reinforce nor cancel each other, the best guess of either innovation will be less than one. Equivalently, with uncorrelated innovations, the fraction of an unforecasted change that is likely to be permanent will be less than one. Under the uncorrelated assumption, a 1 percent increase in GNP today will on average therefore lead to a long run increase of less than 1 percent. In general, then, the initial assumption about this correlation is crucial to ascertaining the relative importance of the two components.

The second pair of measures in Table 1 answers a related question about the relative importance of the trend and stationary components: what fraction of the quarterly variation in GNP is attributable to permanent shifts? Unfortunately, the answer to this is ambiguous, even given an estimated model for GNP, since the fraction of the variance accounted for by the trend depends on the correlation between the trend and stationary innovations. The two measures in the table are "R-squared" statistics that would arise from regressing the quarterly change in GNP against the change in its true trend which in general is unobserved where the trend is computed under different assumptions about this correlation. The first statistic, $R^2_{\text{corr}=1}$, is from the regression using the trend computed assuming the trend and stationary innovations to be perfectly correlated, so that this trend is computed using Beveridge and Nelson's formula. The second measure is based on the observation that a given ARIMA model cannot be distinguished from (formally, is observationally equivalent to) a UC-ARIMA model with an assumed correlation falling in a certain range, where this range depends on the parameters of the ARIMA model. Accordingly, $R^2_{\text{corr}=\text{min}}$ is computed using the smallest correlation that is capable of generating the estimated ARIMA model from a UC-ARIMA model. Thus these R^2 measures provide a rough range of the fraction of movements in quarterly GNP attributable to shifts in the trend.[8]

---

[8] The observational equivalence of ARIMA and UC-ARIMA models for a range of correlations is not obvious, although as an illustration it is readily derived for the UC-ARIMA model, $y_t = y_t^p + e_t^s$, where $y_t^p = y_{t-1}^p + e_t^p$, $\text{cov}(e_t^s, e_t^p) = 0$, and where e_t^p and e_t^s are serially uncorrelated. This model is observationally equivalent to the ARIMA(0, 1, 1) model: $\Delta y_t = e_t + be_{t-1}$, where $b/(1 + b^2) = -\text{var}(e_t^s)/[\text{var}(e_t^p) + 2\text{var}(e_t^s)]$. The observational equivalence follows because both models imply the same autocovariances for Δy_t.

Our variance measures are inspired by Masanao Aoki's (1987) measure of the relative variance of the two components. The correlation for the $R^2_{\text{corr}=\text{min}}$ measure was computed by matching the autocovariances implied by the estimated ARIMA model to those implied by a UC-ARIMA model with various correlations between the two innovations. The R^2 refers to the trend from the UC-ARIMA model with the smallest correlation for which this match was possible, assuming an ARIMA(0, 100) structure for the stationary component in the UC-ARIMA model. The numerical search over trial correlations was carried out using steps of .05. The search was initialized by setting the autocovariances of the stationary process equal to the autocovariances of the innovation in GNP growth.

According to the first, "perfect correlation" measure, all the univariate models attribute much of the quarterly movements in GNP to shifts in its trend. Indeed the theoretical maximum of this R^2 is attained in the perfect correlation case. Examining the range of R^2 measures indicates that although the assumed correlation makes a quantitative difference to the estimated fraction of the variance associated with movements in the trend, in all cases it exceeds one-third, and in all but one it exceeds one-half.[9]

Do these calculations resolve the basic question of the extent to which quarterly movements in GNP are associated with variations in its trend? The answer must be both yes and no. Certainly these results provide further substance to the original findings of time series analysts that aggregate economic variables, and in particular GNP, seem to be integrated processes and therefore contain a stochastic trend. Indeed, even the smallest of the R^2 measures (.34) and of the estimates of the change in the long run forecast (.6) are qualitatively quite different from the values of zero that would obtain were the trend in GNP deterministic rather than stochastic. These statistical measures therefore provide further confirmation of the conclusions drawn from our initial examination of GNP in Figure 1: changes in trends seem to be an important feature of the postwar U.S. experience.

Unfortunately, the ranges of the statistics in Table 1 indicate that the answer must also partially be no. There is nothing intrinsic to this question that rules out the possibility of reaching agreement among the various models. But, at least for the U.S., the different modeling strategies lead to very different quantitative conclusions.[10]

We have emphasized that a key reason for this ambiguity is the choice of the correlation of the innovations in the permanent and transitory components, likening this to the difficulty of deciding whether the 1975 downturn was a shift in the permanent or transitory components based solely on visual examination of GNP in Figure 1. Ultimately, however, as economists we should object to having one hand tied behind our back when sorting out trend and stationary movements, in the sense that this analysis focuses solely on the evidence contained in historical GNP. Rather, we might ideally wish to use qualitative evidence about the behavior of the monetary authorities in response to the quadrupling of oil prices in 1973–1974, or about the lasting effect on investment, technology, and human capital of the adjustments made in response to the price hike. Since such information is difficult to use in the context of statistical models, we still might hope to perform a less ambitious analysis using additional aggregate variables such as employment, consumption and investment.

---

[9] One approach to resolving the ambiguous results in Table 1 is to develop sensitive statistical measures to assist in choosing one model from another. For example, Cochrane suggested examining the variance of the growth in GNP over several years as a means of estimating the importance of stationary fluctuations in determining the long run behavior of GNP; Joseph Gagnon (1986) has taken a different statistical approach to the same question. Both provide evidence that the low order ARMA models overstate the importance of the trend component in GNP, essentially for the reasons discussed above. However, Campbell and Mankiw (1987b) provide evidence that variance ratios involving long differences might exhibit substantial small sample bias when based on quarterly data covering 30–40 years.

[10] These different models might yield similar conclusions when applied to data from different countries. In an initial comparison of the stochastic trend behavior across countries, Clark (1987b) has provided evidence of striking international differences in the relative importance of the trend and stationary components. Indeed, for some countries he suggests that the notion of a stochastic trend as it has been discussed here is inappropriate; following Harvey, he argues for a model in which the trend itself follows a random walk. In this case, GNP would be modeled as being integrated of order two.

Using Additional Information To Study Macroeconomic Trends

Casual inspection of Figure 1 suggests not only that GNP, consumption and investment appear to contain a stochastic trend, but that they contain a common stochastic trend. Indeed, one indication that the "trend" in GNP changed between the 1960s and the 1970s is that this shift appears in consumption and investment as well. This suggests that consumption and investment data contain important information about the trend in GNP, so that there is likely to be a statistical payoff to analyzing these series jointly.[11] This section briefly describes some recent work that uses multiple aggregate variables to assess the importance of the trend component in output.

One device used in these multivariate studies to measure the importance of the innovation to the permanent component is the concept of "variance decompositions," which Christopher Sims (1980a, 1980b) has profitably applied to Vector Autoregressive (VAR) multivariate time series models. When a forecast of GNP (econometric or otherwise) misses its target, and when this forecast error can be traced to a particular factor, it is reasonable to conclude that this factor is important in determining the evolution of GNP—at least within the context of the model that generated the forecast. Forecast error variance decompositions build on this intuition by quantitatively attributing the errors in forecasting the different variables to the various innovations in the system. For example, suppose that, in an empirical multivariate model with permanent and transitory components, forecasts of GNP two years hence are typically off by ±1.5 percent. These forecast errors will, by construction, arise from errors in forecasting the trend, the stationary component, or both. If the errors from forecasting the trend generally exceed the errors from forecasting the transitory component, then one might attribute more importance to the permanent than the transitory component in determining the evolution of GNP two years hence. In multivariate models with stochastic trends, Sims' forecast error variance decomposition performs this calculation.

Several recent studies (Olivier Blanchard and Danny Quah, 1987; Clark, 1987b; Campbell and Mankiw, 1987b; Andrew Harvey and Stock, 1987; Matthew Shapiro and Watson, 1988) have used multiple aggregate variables to shed greater light on the importance of the permanent component in GNP. Blanchard and Quah (1987) and Robert King, Plosser, Stock, and Watson (1987) both use these forecast error variance decompositions as a guide to assessing the importance of the permanent component.

---

[11] Since government expenditures and net exports are not considered, the national income accounting identity for GNP imposes no restrictions on the number of common trends in GNP, consumption and investment.

Blanchard and Quah (1987) identify this component by assuming that it has no permanent effect on unemployment. In contrast, King, Plosser, Stock, and Watson identify the permanent component in output by assuming that it is also the permanent component in consumption and investment. Despite these and other differences, both papers conclude that between 60 percent and 80 percent of the movements in output at the two- to four-year horizon are explained by movements in the permanent component. Although this literature is still developing, an emerging conclusion from these multivariate studies is that permanent innovations in output play an important role in determining the movements of GNP at horizons typically associated with the business cycle.

Implications

There is a large body of evidence that macroeconomic variables behave as if they contain stochastic trends. Moreover, the empirical research outlined here suggests that the innovations in these stochastic trends play an important role in short run cyclical movements. Multivariate empirical analysis suggests that trend variations and business cycle movements appear to be related. One interpretation of this link is that business cycle fluctuations might be caused by innovations in growth. An alternative explanation—equally consistent with the empirical results—is that cyclical fluctuations cause changes in long run growth. This latter view is consistent with James Tobin's (1980) argument, "With respect to human capital, as well as to physical capital, demand management has important long run supply-side effects. A decade of slack labor markets, depriving a generation of young workers of job experience, will damage the human capital stock far beyond the remedial capacity of supply-oriented measures." Given the challenges associated with differentiating the trend and cyclical components in the time series models discussed above, however, the logical next step is to bring additional information to bear on distinguishing the complex trend-cycle interactions with which Tobin was concerned.

Interpreting Econometric Evidence When Variables Have Stochastic Trends

A Tale of Two Econometricians

Consider the plight of two hypothetical econometricians studying aggregate consumption—but ignorant of the pitfalls that can arise when performing econometric analysis with integrated variables. The econometricians do not, of course, know the true structure of the economy which produced their data; but we do. Specifically, we have constructed for them a simple artificial economy, focusing on aggregate real per capita consumption (C_t), aggregate disposable income (Y_t), and a price index (P_t).

This artificial economy has several key features. First, disposable income consists of two parts: in Milton Friedman's (1957) terminology, permanent and transitory income. The permanent component of disposable income is assumed to follow a random walk, while the transitory component is an independently and identically distributed random variable that is independent of the permanent component. Using the terminology of the previous section, we therefore assume that disposable income has a stochastic trend plus a stationary component, that this stationary component has no serial correlation, and that the correlation between the trend and stationary innovations is zero.

Second, we suppose that consumers know their permanent income and that they behave according to a narrow interpretation of Friedman's permanent income hypothesis, thereby consuming precisely the permanent component of their disposable income (which changes from period to period, since it is a random walk). Finally, price changes—assumed to be random and unforecastable with mean zero—do not

Table 2

Regression results using data from the artificial economy (3)–(7)

Typical Regression Results			*Results based on 1000 Replications*			
Estimated Regression Equation	R^2	*D.W.*	*Percent rejections using the usual 10% two-sided t-test*		*Other statistics*	
Econometrician #1						
1. $C_t = \underset{(28.7)}{9.16} + \underset{(5.12)}{.40}\ P_t$	.15	.08	Testing $\beta_{P_t} = 0$:	81%	Median $R^2 =$	.15
2. $C_t = \underset{(6.35)}{2.48} + \underset{(16.9)}{.069}\ t$	.66	.16	Testing $\beta_t = 0$:	91%	Median $R^2 =$	.42
3. $\Delta C_t = \underset{(.81)}{.048} + \underset{(8.06)}{.28}\ \Delta Y_t$	.31	2.27	—		Median $\hat{\beta}_{\Delta Y_t} =$	.33
4. $\Delta Y_t = \underset{(2.33)}{.41} - \underset{(-2.15)}{.041}\ C_{t-1}$	.03	1.98	Testing $\beta_{C_{t-1}} = 0$:	48%	Median $\hat{\beta}_{C_{t-1}} =$	$-.035$
Econometrician #2						
5. $C_t = \underset{(2.60)}{.51} + \underset{(-2.74)}{.94}\ Y_t$	.93	1.88	Testing $\beta_{Y_t} = 1$:	84%	Median $\hat{\beta}_{Y_t} =$	.94
6. $C_t = \underset{(2.52)}{.45} + \underset{(11.7)}{.97}\ C_{t-1} - \underset{(-.13)}{.01}\ C_{t-2}$	.94	2.01	Testing $\beta_{C_{t-2}} = 0$:	9%	Median $\hat{\beta}_{C_{t-2}} =$	$-.001$
7. $C_t = \underset{(2.36)}{.41} + \underset{(14.3)}{1.03}\ C_{t-1} - \underset{(-.97)}{.07}\ Y_{t-1}$	.94	1.97	Testing $\beta_{Y_{t-1}} = 0$:	10%	Median $\hat{\beta}_{Y_{t-1}} =$	.005
8. $C_t = \underset{(2.21)}{.47} + \underset{(45.0)}{.95}\ C_{t-1}$	.94	2.00	Testing $\beta_{P_{t-1}} = 0$:	27%	Median $\hat{\beta}_{P_{t-1}} =$	$-.003$
$+ \underset{(.17)}{.004}\ P_{t-1} + \underset{(.87)}{.06}\ \Delta P_{t-1}$			Testing $\beta_{\Delta P_{t-1}} = 0$:	9%	Median $\hat{\beta}_{\Delta P_{t-1}} =$	$-.008$

Notes: The regression results in the left half of the table are based on a typical draw of 150 observations constructed according to (1)–(4) using a random number generator in the statistical package RATS (see footnote 12). R^2 and D.W. respectively denote the regression R-squared and the Durbin-Watson statistic. t-statistics are given in parentheses. The t-statistic on the coefficient on Y_t in regression 5 tests the hypothesis that the coefficient equals one, while all other t-statistics refer to the hypothesis that the coefficient equals zero. The entries in the right half of the table summarize the results of repeating these regressions 1000 times with independently drawn series.

confuse these consumers, in the sense that real consumption and disposable income are determined independently of the price level or its changes.

To summarize these assumptions mathematically, let Y_t^p and Y_t^s respectively denote the permanent and stationary (or transitory) components of disposable income (Y_t). The equations describing output, consumption and the price level in this artificial economy are:

$$Y_t = Y_t^p + Y_t^s \tag{3}$$

$$Y_t^p = Y_{t-1}^p + u_t \tag{4}$$

$$C_t = Y_t^p \tag{5}$$

$$P_t = P_{t-1} + v_t. \tag{6}$$

The innovations Y_t^s, u_t, and v_t are assumed to be mutually independent and, for convenience, to be normally distributed with mean zero and unit variance. According to (3), disposable income (Y_t) is the sum of its permanent (Y_t^p) and transitory (Y_t^s) components where, according to (4), the permanent component evolves according to a random walk. The consumption function (5) states that consumers set their consumption (C_t) equal to their permanent income, so that the marginal propensity to consume out of permanent income is one. Finally, (6) states that, like permanent income, the price level (P_t) is a random walk.

We provide our two econometricians with a typical time series comprised of 150 observations (coincidentally, the number of quarters between 1950:I and 1987:II) on the three variables Y_t, C_t and P_t, generated according to (3)–(6) using a pseudo-random number generator on a portable computer.[12] They do not know but seek to uncover the true relations among these variables. What might they learn?

The first econometrician begins by investigating whether consumers change their consumption patterns based on the price level, which he does by regressing C_t against a constant and P_t. Since the t-statistic on P_t (given on the left hand side of Table 2) of

---

[12] The following computer program, written for the econometrics package RATS, will produce random time series data according to equations (3)–(6):

```
cal 50 1 4
all 0 87:2
dec rect err(150, 3)
matrix err = ran(1.0)
zer yp ; zer ys ; zer p
eval yp(1) = err(1, 1) ; eval ys(1) = err(1, 2) ; eval p(1) = err(1, 3)
do i = (50:2), (87:2)
   eval yp(i) = yp(i - 1) + err(i, 1)
   eval ys(i) = err(i, 2)
   eval p(i) = p(i - 1) + err(i, 3)
end do i
set y / =  yp(t) + ys(t) ; set c / =  yp(t)
diff y / 1 dy ; diff c / 1 dc ; diff p / 1 dp
```

5.12 far exceeds the conventional 5 percent two-sided critical value of 1.96, a standard interpretation of his regression is that nominal prices affect consumption. He then checks whether consumption has a linear deterministic time trend by regressing C_t on a constant and time; upon checking the t-statistic, he concludes that it does. However, he recognizes that the low Durbin-Watson statistic from these regressions indicates substantial serial correlation in the residuals (e.g. Robert Pindyck and Daniel Rubinfeld, 1981, pp. 158–164). Troubled by these low Durbin-Watson statistics and thinking consumption to have a deterministic time trend, he differences the data and attempts to estimate the marginal propensity to consume by regressing the change of consumption on the change of income; he finds that the marginal propensity is small indeed. Finally, he recalls Robert Hall's (1978) famous argument that Friedman's Permanent Income Hypothesis implies that consumption follows a random walk, so that the first difference of consumption should be unpredictable. Accordingly, he checks whether lagged consumption is a useful predictor of future changes in consumption; based on the t-statistic of -2.15, he rejects the random walk hypothesis at the 5 percent significance level. Summarizing, he concludes that consumers have money illusion, that consumption contains a linear time trend, that the marginal propensity to consume is .28, and that past values of consumption are useful in predicting future consumption. That he drew these conclusions is not an artifact of the particular series we gave him to analyze. Repeating this experiment 1000 times using independent draws (the results are summarized on the right hand side of Table 2) indicate that his findings were typical.

Each of his conclusions is wrong.

The second econometrican estimates different regressions. She estimates the marginal propensity to consume by regressing consumption against income and finds it large, but significantly less than one using the usual 5 percent critical value for the t-statistic. When she tests the random walk hypothesis by regressing consumption on two of its lags, the second lag has no statistically significant predictive content; the same conclusion obtains if a lag of income is used. Finally, she finds no additional forecasting value of lagged price changes (although, with the benefit of 1000 replications, we know she would incorrectly reject the hypothesis of no predictive content of the lagged price level 27 percent of the time using the 10 percent critical value). She concludes that Hall's interpretation of Friedman's theory is valid and that the marginal propensity to consume is less than one, although this latter finding seems to be more a matter of statistical than economic significance.

Her conclusions, then, are largely right. Participants at a conference at which these two econometricians present their results might find the exchange entertaining. But they might also long for a systematic way to decide which regression results could be trusted and which could not.

Recent Developments in the Theory of Regression with Integrated Variables

It has long been recognized that the usual techniques of regression analysis can result in highly misleading conclusions when the variables contain stochastic trends. In the econometrics literature, since Clive Granger and Paul Newbold's (1974) influential simulation study, this has been known as the problem of "spurious regressions." Largely influenced by the techniques of Box and Jenkins, the accepted "solution" to the "problem of nonstationarity" has been to transform the variables so that they appear to be stationary; in practice this typically means using first differences of the series. Unfortunately, by sidestepping the issues raised by stochastic trends, this approach has little to say about the regressions in Table 2. Moreover, simply using first differences of the data in the regressions generally will not suffice to uncover the true relations in the economy, as the first econometrician's regression of changes in consumption on changes in income makes clear.[13]

The past few years have seen important progress associated with the specification and analysis of multivariate models with integrated processes. Much of this stems from Fuller's (1976) and Dickey and Fuller's (1979) development of the first formal tests for the existence of stochastic trends in a single time series, and from three seminal papers by Granger and Andrew Weiss (1982), Granger (1983), and Robert Engle and Granger (1987). These latter papers provide a mathematical framework for analyzing variables that contain common stochastic trends. Specifically, they consider the case that two (or more) variables might each contain a stochastic trend, i.e. appear to be integrated: consumption and GNP in Figure 1 are good examples. However, casual inspection suggests that these series contain a common trend—and that, by subtracting out this trend, the difference between the two variables is stationary. Formally, they define two integrated process to be *cointegrated* if there is some linear combination (that is, weighted average) of them that is stationary. Thus consumption and GNP are arguably individually integrated; indeed, Dickey and Fuller's tests fail to reject this null hypothesis. But, assuming that log consumption less log output is stationary, they are jointly cointegrated; that is, they share a common stochastic trend. These theoretical developments (tests for a series being integrated and the concept of cointegration) spurred an enormous amount of recent research into econometric issues that arise when cointegrated processes contain unit roots.[14]

To understand the regression results in Table 2, we focus on one of the key lessons of this research: in certain circumstances, even if the right-hand variables (the regressors) are integrated, the usual procedures of OLS analysis can still provide a satisfactory framework for evaluating econometric evidence and for producing forecasts. Furthermore, this research has produced some simple rules-of-thumb that suggest when this is—and is not—likely to be the case.

---

[13] In a lively discussion of the history of spurious regression, David Hendry (1986) traces the recognition of this problem in the context of integrated processes to G. Yule (1926). In many circumstances, using differences of time series variables has proven very successful. Aside from having the intuitive appeal of modeling the rates of change of variables (when first differences of logarithms of the series are used), modeling differences of variables is arguably well-suited to producing short run forecasts—as dramatized by the early success of Box-Jenkins time series methods when pitted against the large Keynesian econometric models of the 1960s and 1970s. However, restricting econometric attention to differences in series rules out direct examination of the relation among the levels of the series. Even if the objective is to produce short run forecasts, the econometrician might wish to follow Sims (1980a) and use the additional information available in the levels of variables.

[14] The empirical results of James Davidson, David Hendry, Frank Srba, and S. Yeo (1978) provided an important motivation for the development of the theory of cointegration by developing an empirical model of consumption in which consumption and income were implicitly modeled as cointegrated. Granger (1986) and Hendry review recent developments in this area and discuss the link between error-correction models and cointegration.

We first state these rules-of-thumb in a general context, then apply them to the regressions in Table 2.[15] Broadly speaking, the usual assumptions of time series analysis are:

(i) The error term is serially uncorrelated and is uncorrelated with the regressors (i.e. the regressors are either exogenous or predetermined and the error term is i.i.d.).
(ii) All the regressors are either deterministic or stationary random variables.

Under these circumstances, the estimated coefficients will become arbitrarily close to their true values in increasingly large samples (that is, they will be consistent). Furthermore, in large samples the null distribution of regression *t*- and *F*-statistics can be approximated by normal and *F*-distributions, respectively.

When some or all of the regressors are integrated processes, condition (ii) clearly is violated. Perhaps surprisingly, however, in many cases the usual techniques of regression analysis will still apply. Specifically, suppose that one is interested in interpreting a particular coefficient or set of coefficients in the regression equation. Although (ii) does not hold, suppose that (ii′) does:

(ii′) If there are integrated regressors, either (a) the coefficients of interest are coefficients on mean zero stationary variables; or (b) even if some or all of the coefficients of interest are coefficients on integrated regressors, the regression equation can nevertheless be written in such a way that all the coefficients of interest become coefficients on mean zero stationary variables.

Even if condition (a) in (ii′) does not hold, condition (b) still might; an example of this is given below. Under (i) and (ii′), the OLS estimator will be consistent. Moreover, the *t*- and *F*-statistics for the coefficient(s) of interest have their usual large sample distributions, so (for example) the standard critical values apply.

In some cases it might be impossible to express the coefficient of interest as a coefficient on a mean zero stationary variable, and instead (ii″) might hold:

(ii″) The parameter of interest is a coefficient on an integrated process and cannot be written as a coefficient on a stationary variable.

---

[15] These rules-of-thumb are drawn from Sims, Stock and Watson (1986). For expositional simplicity, we assume throughout that all regressions include a constant term for reasons discussed in that paper. All statements assume that the variables are either stationary or integrated of order one, and that the integrated variables have zero drift. The situation with nonzero drift or multiple orders of integration is somewhat more complicated, and the reader is referred to Sims, Stock and Watson or Stock and Kenneth West (1988) for a discussion of this case.

Under (i) and (ii″), the estimator of the coefficient is consistent, but it does not have the usual normal asymptotic distribution, so the usual critical values do not apply.

When the level of one variable is regressed against the level of another, it might well be the case that the error term is not i.i.d. or independent of the regressors. However, suppose that (i) does not hold, but that (i′) does:

(i′) The integrated dependent variable is cointegrated with at least one of the integrated regressors, so that the error in the regression equation is stationary but not necessarily serially uncorrelated or independent of the regressors.

Under (i′) and (ii′), unless the regressor is strictly exogenous the stationary regressor will typically be correlated with the error term and the parameter estimate will be inconsistent. This is the usual source of "simultaneous equations bias," "omitted variables bias," and "errors-in-variables bias." However, under (i′) and (ii″), *rather* remarkably the estimator of the coefficient of interest is consistent, although it does not have an asymptotic normal distribution.

Finally, suppose that the regressor and at least one dependent variable are integrated, but that there is no cointegrating relationship between the dependent variable and the regressors; it follows that the error term in the regression is integrated. In this case the estimated coefficients on regressors satisfying (ii″) will not be consistent; indeed, these coefficients and the R^2 of the regressions converge to random variables. An important example of this case is the regression of one random walk on another independent random walk, which is Granger and Newbold's (1974) spurious regression problem.

While these rules-of-thumb are rather involved, some intuition for why they work can be developed by first considering the familiar case of a single regressor that is stationary and uncorrelated with a serially uncorrelated error term. Under these assumptions, the OLS estimator is consistent because in large samples the average squared residual formed using any trial coefficient is minimized at or near the true coefficient value; for all coefficients except the true one, the average squared residual is larger but finite. In contrast, in time series regression with two cointegrated variables (so that the dependent variable and the regressor are both integrated but the true error term is stationary), because the two series are trending together, the residual constructed using other than the true coefficient will itself be integrated. Since an integrated process has a variance that tends to infinity, the average squared residual formed using a trial coefficient will grow arbitrarily large as the sample size increases, and will remain finite (and therefore be minimized) only for the true parameter value. This suggests that the coefficient on an integrated regressor can be estimated unusually precisely, as long as the error term is stationary. Furthermore, this unusual behavior of the average squared error suggests that the standard Gaussian asymptotic theory might not apply when there are integrated regressors. In summary, the coefficient in a regression of one integrated variable on another will be consistent if the two variables are cointegrated although, as it turns out, the asymptotic theory is nonstandard.[16]

This intuition can be extended to a regression that includes one stationary regressor, one integrated regressor, and a serially uncorrelated error term that is uncorrelated with either regressor. The preceding reasoning suggests that the coefficient on the integrated regressor will be estimated more precisely than usual (and will have a nonstandard asymptotic distribution); indeed, by the logic of cointegrated regressions, this would be so even if the stationary regressor were omitted from the specification. Turning to the coefficient on the stationary regressor, suppose that the stationary and integrated regressors are uncorrelated. Then the usual reasoning for OLS with uncorrelated regressors suggests that this coefficient will have the conventional large-sample properties. While this argument assumed the two regressors to be uncorrelated, this turns out not to be restrictive: it can be shown that, since the stochastic trend dominates the behavior of an integrated process, the sample correlation between an integrated variable and any mean zero stationary variable tends to zero as the sample size increases. Thus the usual tools of time series regression can be used to examine the coefficient on the mean zero stationary regressor even if there are other regressors that are integrated. Finally, this reasoning can be extended to the case that the coefficient can be written as a coefficient on a mean zero stationary regressor by recognizing that the sum of a stationary variable and an integrated variable is itself integrated, so that the previous logic applies directly to the rearranged regression.

We illustrate these general principles by returning to the curious regressions in Table 2.

Understanding the Two Econometricians' Results

These rules-of-thumb provide a simple framework for explaining the results of our two econometricians, especially since we know the true economic structure that generated their data.

By construction, consumption and the price level are integrated processes that do not share a common stochastic trend, so they are not cointegrated. Thus the first regression clearly falls into the category of a spurious regression. As both Granger and Newbold (1974) and Peter Phillips (1986a) have emphasized, an indication of this situation is the extremely low Durbin-Watson statistic—although this statistic does not provide a formal test for a relation between integrated processes.

---

[16] This argument applies both when the regressor and the dependent variable are cointegrated and when the regressor is a lagged value of the (integrated) dependent variable. In the latter case, the estimated coefficient will converge to one, even if additional lags in the regression have been omitted, so that the error is stationary but serially correlated. Assuming the process to have no drift, in this case the estimated coefficient will have an asymptotic "unit root" distribution as discussed by Fuller (1976) and Dickey and Fuller (1979) which differs dramatically from the usual normal distribution, implying substantial bias towards zero in moderate sample sizes. The estimator converges to this distribution at the rate T rather than $T^{1/2}$ as in the case of conventional time series regression. Also see Fuller's discussion of the case that the variable contains a drift.

The second regression also fails to meet any of the criteria for applying the usual asymptotic approximation to the t-statistic, since consumption is not cointegrated with a linear time trend; indeed, from the definition of cointegration, it cannot be, since a linear time trend is not itself an integrated stochastic process. This regression was analyzed theoretically in an influential article by Nelson and Heejoon Kang (1981); in addition to producing a random R^2, with 100–150 quarterly observations the residuals from this regression seem to exhibit cycles with a period similar to one associated with the business cycle. Thus a linearly detrended random walk is likely to exhibit spurious periodicity.

Had our econometrican read Friedman closely, he would have recognized the problem with the third specification: the change in disposable income measures the change in permanent income, but with error, since it includes the change in transitory income as well. Thus the coefficient estimator in the third regression is biased downwards, even in arbitrarily large samples.

His final regression satisfies condition (i) (the true coefficient is zero and the error term is the i.i.d. error in permanent income); however, the regressor C_{t-1} is integrated. Thus the coefficient on C_{t-1} satisfies (ii″). It follows that this coefficient is consistent (indeed, its median estimate is close to its true value of zero), but that it will not have the usual asymptotic distribution. For the hypothetical economy (3)–(6), the t-statistic on C_{t-1} is in fact Dickey and Fuller's proposed test for a unit root in consumption; according to Fuller (1976, Table 8.5.2, p. 373) the correct 10 percent critical value for a test against the hypothesis that consumption is stationary is -2.57. Had the first econometrician used this critical value, he would have failed to reject the hypothesis that consumption follows a random walk.

Why did the second econometrician fare better? By construction, consumption and income are cointegrated in the artificial economy, so the theoretical rules-of-thumb indicate that her first regression will result in a consistent estimator with a nonstandard distribution. Indeed, the median estimate of the 1000 replications is close to—but less than—one, although the t-statistic clearly has a nonstandard distribution. This is a specific example of Stock's (1987) result about the consistency of OLS estimators of cointegrating vectors, even if the error in the regression equation is serially correlated or not independent of the regressors. Additionally, a moment's reflection will indicate that Trygve Haavelmo's (1943) argument that the OLS estimator of the MPC is consistent because of simultaneous equations bias does not hold when consumption and income are cointegrated; rather, the estimator will be consistent, but will exhibit small sample bias.[17]

Her next test involved a coefficient on an integrated process in a regression equation with an i.i.d. error. However, the coefficient on the second lag of consumption can be rewritten as a coefficient on a stationary variable, so that condition (ii′) applies. To see this, consider that part of the regression involving the lags of consumption, and denote the two coefficients as α and β. Algebraic manipulations show that $\alpha C_{t-1} + \beta C_{t-2} = (\alpha + \beta)C_{t-1} - \beta(C_{t-1} - C_{t-2})$; but $C_{t-1} - C_{t-2}$ is stationary, so β clearly can be written as a coefficient on a stationary variable. Thus the theory predicts that the usual t- and F-distributions will apply, and the simulation results on the right side of Table 2 support this prediction. This argument applies equally to her next regression, except that the stationary combination of regressors is $Y_{t-1} - C_{t-1}$, which is stationary because consumption and income are cointegrated.

---

[17] It is interesting to note that this small sample bias provides an explanation of the difference between estimates of the marginal and average propensities to consume that provided one of the original motivations for Friedman's study of consumption. Specifically, applying nonstandard distribution theory to Friedman's original data, Stock (1986) shows that the marginal propensity to consume has a negative bias of approximately .15; upon adjusting for bias, the marginal and average consumption propensities are both approximately .9.

Her final regression satisfies (i) and, since ΔP_{t-1} is stationary, its coefficient satisfies (ii′), so that the usual asymptotic theory applies to this coefficient (and indeed appears to work well in moderately sized samples, at least according to these calculations). However, the coefficient on P_{t-1} cannot be written as a coefficient on a mean zero stationary regressor, since there are no other regressors with which P_{t-1} is cointegrated. It follows that this estimator is consistent (converging to zero in this case) but has a nonstandard distribution, so that the usual critical values do not apply.[18]

These rules of thumb thus provide a general framework for evaluating these regression results with integrated variables, assuming the true economic structure to be known.[19]

Evaluating Actual Regression Results

The previous discussion emphasizes the importance of uncovering the stochastic trend properties of actual data to be used in regression analysis. In certain circumstances, economic theory might suggest orders of integration and cointegration among the variables. For example, Campbell (1987) argues that consumption and income being cointegrated is plausable both theoretically and empirically. Often, however, economic theory provides no clear guidance in determining which variables have stochastic trends, which do not, and when the trends are common among those that do.

There are no simple "recipes" for performing time series analysis with integrated variables. One sensible starting point in analyzing time series data in which there might be stochastic trends, however, is to perform a series of initial tests on the data. In particular, Dickey and Fuller's test for a unit root in a single series can provide an important piece of evidence about whether the variable is integrated. Typically an analysis will involve multiple time series, in which case it is important to know their cointegration properties as well. Diagnostic measures for the existence of cointegration have been proposed and analyzed by Engle and Granger, Søren Johansen (1987), Phillips and Ouliaris (1987), and Stock and Watson (1986).[20] While these tests certainly are not foolproof, their judicious application can shed considerable light on the way that stochastic trends enter the time series. This in turn provides a framework for implementing the rules of thumb described in this section.

---

[18] Interpreting P_t not as the price level but as the Standard and Poor's Stock Price Index, the reader might recognize her final three equations as simplified versions of Hall's tests of the Permanent Income Hypothesis. Stock and West provide further evidence on the finite sample performance of the asymptotic distribution theory that arises in interpreting Hall's regressions.

[19] This discussion has focused on the case of roots exactly equaling one, rather than (say) .999 or 1.001. Mathematically, however, similar warnings about the use of usual regression techniques arise for roots close to one; for example, see Cavanagh (1986) or Phillips (1986b). Thus the rules-of-thumb for identifying those coefficients with asymptotic normal distributions can be seen as tools to guard against misleading inferences if one suspects there to be roots close to one.

Conclusions

Macroeconomic time series appear to contain variable trends. Moreover, modeling these variable trends as random walks with drift seems to provide a good approximation to the long run behavior of many aggregate economic variables, at least in the U.S. While this general observation is over thirty years old, the application of recently developed statistical "magnifying glasses" has led to several important conclusions for developing and testing macroeconomic theories and for formulating macroeconomic policies.

First, variations in growth trends constitute a quantitatively large part of the movements in real per capita GNP in the United States. Thus the importance of shifts in long run prospects must be recognized even if one is primarily concerned with a relatively short specific historical episode.

Second, the presence of stochastic trends requires careful thought to avoid important econometric pitfalls. If an econometrician wishes to exploit the additional information contained in levels of variables rather than their differences, it is possible to apply a variety of tests and some simple rules-of-thumb to reduce the possibility of making dramatic errors in inference.

Finally, there is evidence not only that aggregate variables contain a substantial stochastic trend component, but that there is a link between changes in this stochastic trend and business cycle movements. This emphasizes the importance of assessing both the short run implications of growth policies and the long run implications of stabilization policies.

■ *The authors thank Francis Bator, Peter Clark, Francis Diebold, Eric Fisher, Joseph Gagnon, Deborah Haas-Wilson, Andrew Harvey, Charles Manski, Danny Quah, Carl Shapiro, Joseph Stiglitz, and Timothy Taylor for helpful comments.*

---

[20] Nelson and Plosser provide a clear discussion of the particulars of applying the Dickey-Fuller test to a single series. One shortcoming of integration or cointegration tests is that, with sample sizes typically encountered in macroeconomic research, they can have a fairly low ability to discriminate between the various hypotheses, particularly with multiple variables. For example, with 150 observations it is difficult to distinguish an integrated process from one that is stationary but highly serially correlated. (One implication of this is that tests based on theoretical cointegrating vectors, if they are known, typically have greater power than tests based on estimated cointegrating vectors.) This is closely related to the statistical difficulties discussed in the previous section with distinguishing stochastic trends from stationary components, since both econometric enterprises involve extracting information about relations among variables over the very long run using only 30 to 40 years of data.

Appendix
Stochastic Trends and Integrated Processes

This appendix presents the mathematical link between ARIMA models and the concept of stochastic trends used in the text.

Beveridge and Nelson show that any ARIMA model can be represented as a stochastic trend plus a stationary component, where a stochastic trend is defined to be a random walk, possibly with drift. This representation is most easily obtained for an ARIMA(0, 1, 1) model. Specifically, suppose that Δy_t (where $\Delta y_t = y_t - y_{t-1}$) is a MA(1) process, so that $\Delta y_t = e_t + be_{t-1}$, where e_t is i.i.d. and b is a constant. Let $y_0 = e_0 = 0$, so that y_t can be written as

$$\text{(A.1)} \qquad \begin{aligned} y_t &= y_{t-1} + e_t + be_{t-1} \\ &= y_{t-2} + (e_{t-1} + be_{t-2}) + (e_t + be_{t-1}) \\ &= \sum_{r=1}^{t} e_r + b\sum_{r=1}^{t-1} e_r \\ &= (1+b)\left(\sum_{r=1}^{t} e_r\right) - be_t. \end{aligned}$$

Letting $y_t^p = (1+b)\Sigma_{r=1}^{t} e_r$ and $y_t^s = -be_t$, one can rewrite the final expression in (A.1) as

$$\text{(A.2)} \qquad y_t = y_t^p + y_t^s, \quad \text{where } y_t^p = y_{t-1}^p + (1+b)e_t.$$

Evidently y_t^p is a random walk with no drift and y_t^s is stationary (indeed, here y_t^s is serially uncorrelated). Equation (A.2) gives the Beveridge-Nelson representation of an ARIMA(0, 1, 1) process in terms of a stochastic trend (y_t^p) and a stationary component (y_t^s). Note that the innovations in the two components are both proportional to e_t, i.e. they are perfectly correlated; if $b > 0$, the correlation is -1, whereas if $b < 0$, the correlation is $+1$.

The representation can in fact be obtained for general ARIMA(p, 1, q) processes. This is most easily shown using lag (or "backshift") operator notation, where by definition $L^j x_t = x_{t-j}$ and $a(L) = \Sigma_{j=0}^{p} a_j L^j$. With this notation, an ARIMA(p, 1, q) model can be written,

$$\text{(A.3)} \qquad a(L)\Delta y_t = f + b(L)e_t$$

where f is a constant and $a(L)$ and $b(L)$ are lag polynomials of order p and q, respectively. Inverting $a(L)$, one can write (A.3) in its infinite moving average form,

$$\Delta y_t = g + c(L)e_t \tag{A.4}$$

where $g = f/\Sigma_{j=0}^{p} a_j$ and $c(L) = b(L)/a(L)$. Next, recursively substitute lagged Δy_t as was done in (A.1) and assume that $y_0 = 0$ and $e_r = 0$ for $r \leq 0$, so that an expression similar to the final one in (A.1) obtains:

$$y_t = gt + h\sum_{r=1}^{t} e_r + d(L)e_t \tag{A.5}$$

where $h = \Sigma_{j=0}^{\infty} c_j$ and $d_i = -\Sigma_{j=i+1}^{\infty} c_j$. Thus (A.5) can be rewritten as

$$y_t = y_t^p + y_t^s, \quad \text{where} \quad y_t^p = g + y_{t-1}^p + he_t \quad \text{and} \quad y_t^s = d(L)e_t \tag{A.6}$$

which provides the Beveridge-Nelson decomposition for general ARIMA($p, 1, q$) processes, where the stochastic trend has drift g. Note that h is easily calculated as $\Sigma_{j=0}^{q} b_j/\Sigma_{i=0}^{p} a_i$. As in the ARIMA(0, 1, 1) case, the innovations in the trend and the cyclical components are both proportional to e_t, and thus are perfectly correlated.

References

Aoki, Masanao, "How to Build State Space Models for Nonstationary Time Series and How to Measure Random Walk Components," University of California-Los Angeles, Department of Economics Working Paper No. 438, March 1987.

Beveridge, Stephen, and Charles R. Nelson "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics*, 1981, *7*, 151–174.

Blanchard, Olivier J., and Danny Quah, "The Dynamic Effects of Aggregate Demand and Supply Disturbances," manuscript, MIT, 1987.

Box, George E. P., and Gwilym M. Jenkins, *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day, 1970.

Campbell, John Y., "Does Saving Anticipate Declining Labor Income? An Alternative Test of the Permanent Income Hypothesis," *Econometrica*, 1987, *55*, 1249–1274.

Campbell, John Y., and N. Gregory Mankiw (1987a), "Are Output Fluctuations Transitory," *The Quarterly Journal of Economics*, 1987, *102*, 857–880.

Campbell, John Y., and N. Gregory Mankiw (1987b), "Permanent and Transitory Components in Macroeconomic Fluctuations," *The American Economic Review*, 1987, *77*, 111–117.

Cavanagh, Christopher L., "Roots Local to Unity," Harvard University Department of Economics Discussion Paper, 1986.

Clark, Peter K. (1987a), "The Cyclical Component of U.S. Economic Activity," *The Quarterly Journal of Economics*, 1987, *102*, 797–814.

Clark, Peter K. (1987b), "Trend Reversion in Real Output and Unemployment," Discussion Paper, Graduate School of Business, Stanford University, 1987.

Cochrane, John H., "How Big is the Random Walk Component in GNP," *Journal of Political Economy*, forthcoming.

Davidson, James E. H., David F. Hendry, Frank Srba, and S. Yeo, "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumer's Expenditure and Income in the United Kingdom," *Economic Journal* 1978, *86*.

Dickey, David A. and Wayne A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 1979, *74*, 1057–1072.

Engle, Robert F., and Clive W. J. Granger, "Co-Integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, 1987, *55*, 251–276.

Friedman, Milton, *A Theory of the Consumption Function*. Princeton: Princeton University Press, 1957.

Fuller, Wayne A., *Introduction to Statistical Time Series*. New York: Wiley, 1976.

Gagnon, Joseph E., "Short Run Models and Long Run Forecasts," *Quarterly Journal of Economics*, forthcoming.

Granger, Clive W. J., "Co-integrated and Error Correcting Models," UCSD Discussion Paper no. 83-13a, 1983.

Granger, Clive W. J., "Developments in the Study of Cointegrated Economic Variables," *Oxford Bulletin of Economics and Statistics*, August 1986, *48*, No. 3, 213–228.

Granger, Clive W. J., and Andrew A. Weiss, "Time Series Analysis of Error-Correction Models," UCSD Discussion Paper no. 82-28, 1982.

Granger, Clive W. J., and Paul Newbold, "Spurious Regressions in Econometrics," *Journal of Economics*, 1974 *2*, 111–120.

Granger, Clive W. J., and Paul Newbold, *Forecasting Economic Time Series*. New York: Academic Press, 1977.

Haavelmo, Trygve, "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 1943 *11*, no. 1, 1–12.

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, 1978, *86*, no. 6, 971–87.

Harvey, Andrew C., "Trends and Cycles in Macroeconomic Time Series," *Journal of Business and Economic Statistics*, 1985, *3*, 216–227.

Harvey, Andrew C., *Time Series Models*. New York: Wiley, 1981.

Harvey, Andrew C. and James H. Stock, "Continuous Time Autoregressive Models with Common Stochastic Trends," manuscript, Harvard University, 1987; forthcoming, *Journal of Economic Dynamics and Control*.

Hendry, David F., "Econometric Modelling with Cointegrated Variables: An Overview," *Oxford Bulletin of Economics and Statistics*, August 1986, *48*, No. 3, 201–212.

Johansen, Søren, "Statistical Analysis of Cointegrating Vectors," manuscript, University of Copenhagen, 1987; forthcoming, *Journal of Economic Dynamics and Control*.

King, Robert G., Charles I. Plosser, James H. Stock, and Mark W. Watson, "Stochastic Trends and Economic Fluctuations," NBER Discussion Paper No. 2229, 1987.

Nelson, Charles R., and Heejoon Kang, "Spurious Periodicity in Inappropriately Detrended Time Series," *Econometrica*, 1981, *49*, 741–751.

Nelson, Charles R., and Charles I. Plosser, "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 1982, 129–162.

Orcutt, Guy H., "A Study of the Autoregressive Nature of the Time Series Used for Tindbergen's Model of the Economic System of the United States, 1919–1932," *Journal of the Royal Statistical Society B*, 1948, *10*, no. 1, 1–45.

Phillips, Peter C. B. (1986a), "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 1986, *33*, 311–340.

Phillips, Peter C. B. (1986b), "Towards a Unified Asymptotic Theory of Autoregression," Yale University, Cowles Foundation Discussion Paper no. 782, 1986.

Phillips, Peter C. B., and S. Ouliaris, "Asymptotic Properties of Residual Based Tests for Cointegration," Yale University, Cowles Foundations Discussion Paper no. 347 (1987).

Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, second edition. New York: McGraw-Hill, 1981.

Prescott, Edward C., "Theory ahead of Business Cycle Measurement," Carnegie-Rochester Conference on Public Policy, 1987.

Shapiro, Matthew D., and Mark W. Watson, "Sources of Business Cycle Fluctuations,"*NBER Macroeconomics Annual*, 1988, forthcoming.

Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48* (1980a), 1–48.

Sims, Christopher A., "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review*, May 1980, *70* (1980b), 250–257.

Sims, Christopher A., James H. Stock, and Mark W. Watson, "Inference in Linear Time Series Models with Some Unit Roots," manuscript, Stanford University, 1986.

Stock, James H., "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, 1987, *55*, 1035–1056.

Stock, James H., "A Reexamination of Friedman's Consumption Puzzle," *Journal of Business and Economic Statistics*, forthcoming.

Stock, James H., and Mark W. Watson, "Testing for Common Trends," *Journal of the American Statistical Association*, forthcoming.

Stock, James H., and Kenneth D. West, "Integrated Regressors and Tests of the Permanent Income Hypothesis," *Journal of Monetary Economics*, 1988, in press.

Tobin, James, "Stabilization Policy Ten Years After," *Brookings Papers on Economic Activity*, 1980, *1*, 19–90.

Watson, Mark W., "Univariate Detrending Methods with Stochastic Trends," *Journal of Monetary Economics*, 1986, *18*, 49–75.

Yule, G. U., "Why Do We Sometimes Get Nonsense-Correlations Between Time-Series?" *Journal of the Royal Statistical Society B*, 1926, *89*, 1–64.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 175–182

Puzzles
Sylvia, Ice Cream and more

Joseph Farrell

In this installment of Puzzles, answers to all puzzles appear in this issue, following the puzzles. As always, readers are invited to criticize the puzzles and answers and to offer new puzzles. Please send all correspondence to: Barry Nalebuff, "Puzzles," c/o *Journal of Economic Perspectives*, Department of Economics, Princeton University, Princeton, New Jersey, 08544. Good luck.

Puzzle 1: Competition as a Welfare Minimum

My teaching assistant Henry was in tears. "It's Sylvia again," he complained. Sylvia is the bane of my class—too clever to take anything for granted, she still gets things spectacularly wrong herself.

"What's she done now?" I asked, wearily. "Well," said Henry, "It all started with consumer and producer surplus. Since we'd already covered the welfare theorem, I thought I'd do a simple calculation to show why competitive equilibrium is optimal in terms of surplus. So I explained how consumer surplus $C(p)$ and total profits $\Pi(p)$ are functions of the price p. Total welfare $W(p)$ is the sum, $C(p) + \Pi(p)$. Now at the welfare-maximizing price p^*, the first-order condition says that $C'(p^*) + \Pi'(p^*) = 0$."

"Yes," I said, "But what does that tell you?" Henry smiled. "Well, this was the clever part—I thought. By the envelope theorem, the derivative $C'(p)$ of consumer surplus with respect to price is just minus the demand, $-D(p)$. And also by the envelope theorem the derivative $\Pi'(p)$ of profits with respect to price is just the competitive supply, $S(p)$. So the first-order condition becomes $-D(p) + S(p) = 0$, which is satisfied for the competitive equilibrium (market-clearing) price but for no other price. I thought this was pretty neat—not really a new proof of the welfare theorem, but at least a good illustrative calculation." He paused and watched my face.

■ *Joseph Farrell is Visiting Assistant Professor of Economics, University of California, Berkeley, California.*

I sucked on my pipe, buying time. Everything seemed to fit together well—too well...? I wondered about the envelope theorem, but I didn't think Henry was bluffing. "Okay," I said, "What next?"

Henry shrank slightly into the special uncomfortable chair I keep for students (so they don't stay too long). "She asked about the second-order condition," he whispered. I relaxed—a mistake when dealing with Sylvia, even at a distance. "Well," I said, a little pompously, "I suppose you just made a concavity assumption?" Henry nodded. "I tried that, of course," he said. "But she said that the second derivative of welfare $W(p)$ is $W''(p) = -D'(p) + S'(p)$, which is always positive (except possibly for Giffen goods). So the competitive-equilibrium price is a *minimum* of welfare, not a maximum!" Henry leaned forward and lowered his voice below a whisper. "Professor Bogart, do you think she's right? Have economists been completely wrong about welfare for decades, just because we didn't check the second-order condition? Think of the liability suits."

I reassured him—economists' advice had never been followed, so there would be no damages. But I was scared. Late that night I came back to my office by the back door, and got out my old grad-school notes. I went over them twice; I didn't know what was wrong. Do you?

•

Puzzle 2: Externalities Galore

In the perfectly competitive economy of Bali Ha'i, there is no production. The natives trade competitively every sundown in picturesque Walras Lagoon with its fine coral reef and its complete markets. The first Western explorer, Captain Smith, came two centuries ago in his ship the *Wealth of Nations* and assured the natives that theirs was the best of all possible worlds—or at least that it could not be made better for everyone. Later, other explorers came and repeated the same message, though written in different ways. Most of the natives just smiled and went back to the beach, but a few studied the new writings. There was a missionary in an isolated study a few miles up the coast, to whose "office hours" they went when they were puzzled.

One day, a man called Sea-Lvia (a distant relative of Sylvia in Problem 1), gave a free lunch in the pineapple grove. After dessert he got to his feet. "Fellow Consumers," he said, his face grave, "We have long believed that our economy is ideal, and the missionaries have told us so. But consider this sad tale.

"Last night at Walras Lagoon I was trying to decide whether to bid for four or for five coconuts, at a price that our friend the Auctioneer here was calling out." (Here some carousers picked up the Auctioneer and carried him around the grove, cheering raucously.) Sea-Lvia continued.

"My friends, that fifth coconut was exactly marginal! And I started thinking how we should choose when we are indifferent. Some would have us toss a fair shell, and thus randomize. Others say it makes no difference what we do. Others ask us to tell the Auctioneer we are indifferent, and let him choose from our demand correspondence; that helps ensure the existence of equilibrium, and our friend is grateful, since if he cannot find an equilibrium he must be sacrificed.

"But, though I am a rationally selfish consumer, I wondered how my choice would affect others: you, my friends. Suppose I said 'five' rather than 'four'. Surely—or almost surely—it would increase the equilibrium price of coconuts. I started thinking how this would affect my friends who buy coconuts—naturally, I ignored the effect on myself, since even thinking about exploiting monopoly or monopsony power is completely tabu, and none of us would ever do it. But is it forbidden to think about the effects on others? Surely not.

"Friends, I was about to say 'four' when I remembered general equilibrium effects. Not only the price of coconuts would be affected! Every price in our economy would be affected! I looked around at my smiling friends, my relatives, the friends of my relatives, the relatives of my friends, the friends of the relatives of my friends..." Some of the natives grew restless: Sea-Lvia was notoriously long-winded. They tossed some papayas at him; he got the message and went on.

"I could not trade last night. Everywhere I turned, there were welfare effects. If the price of honey were to fall, my sister's husband's friend the beekeeper would suffer. If the price of canoe-repairs were to rise, then the already heavy insurance premiums that my uncle pays (for he is not a good driver) would go up. I thought, are not these effects called 'externalities' in other lands? And we are exhorted, if externalities arise, not to ignore them. I could not ignore them. But how could I know what to do? You saw me leave Walras Lagoon. I ate my endowment that night, for I could not conscientiously trade without knowing all the price effects and their welfare consequences."

There was silence in the grove. Certainly it was not tabu to consume one's endowment, though it had never been done before. What was shocking was these externalities. Why had they never noticed them? Hadn't the missionaries assured them that their perfectly competitive economy had none? Yet the argument seemed cogent. And everyone—for they were natural economists—could see that, if there were externalities, then the tribal council should try to make each consumer adjust his demands until his private marginal disutility of adjustment just balanced the marginal external effect. This would put an end to free private optimization. Even in their simple economy, several dozen corrective instruments—some had heard them called Pigovian taxes—would be needed. Someone would have to give up a day's surfing to calculate them; they would need an authority to collect them...the prospect was more appalling than appealing.

People stared at one another, puzzled, worried, and angry. There was no trade that night. Men walked from hut to hut, muttering, conferring. Their faces grew dark. Finally, a youth entered the Public Goods Hut and began to beat a drum. The sound had not been heard in generations, but everyone knew what it meant. It demanded information—or a human sacrifice. They cry went up, "The missionary! Find the missionary!" Nobody suggested waiting for office hours.

You, gentle reader, are the missionary. If you can explain the error in Sea-Lvia's argument, in less time than it takes to set up a human sacrifice (it usually takes about ten minutes), then you live. Otherwise, you become another victim of welfare economics. Get a student to beat a drum and yell, and another to wave a club above your head, while you explain.

Puzzle 3: The Utility of Ice Cream

My mother once told me, when I was about to try, that I should never eat ice cream until I'd had enough, because then I won't have enjoyed it as much as if I had rationed myself. Years later, I formalized this as follows: Suppose my childish utility function, which I would maximize if mom weren't there, is $u(\mathbf{x}) = \Sigma u_i(x_i)$, where x_i is the quantity of good i that I consume. Then my "maternalistic" welfare function, which she will advise me to maximize, is $W(\mathbf{x}) = \Sigma x_i u_i'(x_i)$. Her theory is that enjoyment is largely retrospective, and that how much one savors *each* remembered mouthful of ice cream depends on how fresh and wonderful the *marginal* mouthful tasted—which seems reasonable. So if I consume x_i until its marginal utility is low, I may get less *total* enjoyment out of it than if I restrain myself.

Under what conditions of the utility function $u(\mathbf{x})$ will my mother and I always agree on the amount I should consume, given prices and a budget? Beyond that, when will she want me to consume less of a certain good x_i—say, ice cream—than I would without her advice?

Puzzle 4: Congestion Pricing

Consider a road with a single bottleneck where traffic jams regularly last for two hours in the morning and evening; at other times, there is no congestion. Assuming that the road has a single lane in each direction, and that each driver's time is worth ten dollars an hour, what is the optimal toll to force drivers to internalize the congestion externality, and how does it vary depending on the time of day?

Puzzle 5: Are There Pareto Improvements?

Think of a feasible change in the U.S. economy that does not involve payment of compensation and that would be a Pareto improvement.

Answers

Answer to Puzzle 1

Competitive firms' profits are indeed a convex increasing function of their output price, since the profit function is the upper envelope of straight lines (for each feasible input-output combination, profits are linearly increasing in output price). Similarly, consumer surplus is convex (and decreasing) in any price (except for Giffen goods), since the compensating variation for a price increase dp_i in good i is just $x_i dp_i$, where x_i is the consumer's purchase of good i (this is Roy's identity, or the envelope theorem). So doesn't it follow that any price is better than the competitive price?—or that random prices are better than stable prices?

No it doesn't. If a price other than the competitive equilibrium price is established, then by definition the market does not clear. This means that somebody must be rationed, and does not get the competitive consumer or producer surplus that Henry assumed—people would be unable to trade! Henry's basic formula $W(p) = C(p) + \Pi(p)$ assumes that everyone can trade as much as he wants, and this is not true except at equilibrium. (And, yes, competitive equilibrium does produce a Pareto optimum.) These issues were the subject of some lively debate a while back; see for instance Paul Samuelson (1972).

Incidentally, if you do a similar calculation with quantity instead of price as the "independent variable," you get a sensible answer. The first derivative of welfare with respect to the quantity x that firms produce (at least cost) and consumers consume, is (by the envelope theorem) $-m(x) + d(x)$, where $m(.)$ is marginal cost and $d(.)$ is the demand price. When x^* satisfies the first-order condition, marginal cost must be equal to demand price (marginal value): competitive equilibrium. And the second-order condition just says that the competitive supply curve must meet the demand curve from below.

Answer to Puzzle 2

Sea-Lvia's error is an old one: he noticed *pecuniary externalities*, and thought they were "real" externalities. To see why the external effects of competitive price changes have no welfare consequences, recall that the welfare effect of a price increase dp on a person buying a quantity x of coconuts is just $-x.dp$ in dollar terms. The relevant quantity x is his net purchase, not his consumption. But in the coconut market as a whole, purchases and their opposite, sales, must exactly balance! Therefore, in terms of compensating variation, there is no net welfare effect from a small change in the price of coconuts; and the same goes for every other price change that Sea-Lvia's decision causes. Of course, these price changes have redistributive effects—that is what our friend noticed—but they do not lead to Pareto inefficiency.

If one is concerned with the distributive effects of prices (for instance, if one felt that buyers of coconuts were more worthy than sellers) then some tax or subsidy might be justified. But it would be at the cost of efficiency. In fact, we can relate this to a familiar idea by noticing that if Sea-Lvia were to break the tabu and consider the effect of the price change on his own welfare as a coconut buyer (ignoring the welfare of sellers) he would be acting as a monopsonist. Similarly, if sellers took into account the effects of quantity changes on the well-being of sellers as a group (through changes in price) they would be acting as a cartel.

Answer to Puzzle 3

A useful concept here is the elasticity of marginal utility, which measures how the marginal utility of a good changes with respect to changes in the quantity consumed. Formally, the elasticity of marginal utility (for good i) is the percentage by which the marginal utility of good i falls when the consumption of good i increases by 1 percent.

For a good with a high elasticity of marginal utility, a small increase in the amount consumed will cause a large decrease in marginal utility. On my mother's theory, it makes sense to substitute away from such goods.

An easy *sufficient* condition for my mom and me to agree is that each component utility function $u_i(\cdot)$ has the same constant elasticity of marginal utility, ϵ; that is, each has the form $u_i(x) = a_i x^{\epsilon+1}$ for some constant a_i. Then $x_i u_i'(x_i)$ is a constant (actually, $\epsilon + 1$) times $u_i(x_i)$, and the two overall utility functions—mine and hers—are the same up to a constant; thus they represent the same preferences.

Actually, this sufficient condition is also necessary which we show as follows. Mom and I will always agree on what I should consume if and only if the marginal rates of substitution between any two goods, x_i and x_j, in the functions $u(\mathbf{x})$ and $W(\mathbf{x})$ are equal for all consumption bundles $\mathbf{x}$. Recall that the marginal rate of substitution between goods i and j for a utility function $v(\mathbf{x})$ is the ratio $(\partial v/\partial x_i)/(\partial v/\partial x_j)$. Applying this, we find that the condition for mom and me to agree is that, for all i, j and for all x_i, x_j,

$$\frac{u_i'(x_i)}{u_j'(x_j)} = \frac{x_i u_i''(x_i) + u_i'(x_i)}{x_j u_j''(x_j) + u_j'(x_j)}.$$

This means that $(x_i u_i''(x_i)/u_i'(x_i)) + 1$ must be the same—say, $\epsilon + 1$—for all i and all x_i, which means that $u_i(x_i)$ has the form $k_i x_i^{\epsilon+1}$ for some constant ϵ independent of i; the component utility functions u_i must all have the same constant elasticity. Perhaps the unlikeliness of this condition helps explain why parents and children often argue over ice cream.

If a good i has a higher (absolute) elasticity of marginal utility than the others, that is, $u_i'(x_i)$ falls off more sharply with x_i than is the case for other goods, j, then similar calculations show that in order to maximize $W(\mathbf{x})$ one should cut back on good i relative to the optimal consumption for maximizing $u(\mathbf{x})$. My mother was telling me that ice cream has a high (absolute) elasticity of marginal utility, even for a six-year-old. That is, as one eats more ice cream, its marginal utility falls off faster then is the case for other goods—such as spinach.

As a more challenging exercise in consumer theory, note that the elasticity of marginal utility depends on a particular utility function, not just on preferences (which are the same for two different utility functions, if one is an increasing transformation of the other). Can you translate the condition we have derived for the utility function $u(\cdot)$ into a condition on preferences?

Answer to Puzzle 4

You might think you have too little information, since I haven't told you the number of drivers delayed by a particular driver, nor how long each is delayed. But in fact, you don't need all this information, and you have what you need to know.

Consider the delays I impose on drivers coming after me. Clearly those waiting behind me as I go through the bottleneck are delayed by the length of time, t, that it takes me to do so. But, less obviously, so are all the drivers who arrive after I leave but before the jam clears. They have to wait until the drivers I directly delay have gone through, and so I directly or indirectly delay all the rush hour drivers who come behind me, until the traffic clears.

So I delay each subsequent rush hour driver by t. How many of them are there? Well, suppose it takes a total time T for them all to pass through the bottleneck—that is the time remaining in the rush hour (before the bottleneck next clears). Assuming there's nothing special about me, it presumably takes t for each of them to do so, as it does for me. Then there are T/t of them. The total delay I impose, therefore, is T/t times t, or T, independent of t and of how many drivers are visibly waiting behind me as I reach the bottleneck.

If I travel at the beginning of the rush hour, therefore, I impose a total delay of two hours; and the delay I impose declines linearly down to nothing if I travel at the end of the rush hour. So, if the toll is to internalize that cost, it should be $20 at the beginning of the rush hour, and decline linearly to zero at the end. Of course, it should be zero outside the rush hour, when I impose no delay cost on others.

Three things seem striking about this result. First, it's surprising how little one needs to know in order to calculate the total delay imposed. Second, it's a little counterintuitive that the toll should be highest at the beginning of the rush hour when traffic is probably not all that bad, and not at the peak congestion time. Third, it may surprise some people how large the toll should be. Twenty dollars may seem high; but in some places, such as the San Francisco-Oakland Bay Bridge, a toll collector to whom I spoke insists that it can be all day before the line fully clears! (Remember, it doesn't help if the line gets fairly short later; we only stop counting when the line disappears completely.) Given that we are valuing time at ten dollars an hour, this would mean a three-figure toll. Of course, congestion would fall quite a bit if such a toll were imposed, so that it would be excessive; but it does seem to measure the externalities imposed at present.

Answer to Puzzle 5

I don't have an answer to this one. Perhaps if there were such a change it would be made right away, so that there can never be an answer (just as there can't be a $20 bill on the sidewalk). More likely, the difficulty of the problem reflects the fact that every change affects people in many different ways.

No economist needs to be reminded that tariff reductions are often a good idea—but nor do we need reminding that people fiercely oppose such reductions. We all have our favorite potential improvements in the way the world works, but I have yet to meet any that would hurt nobody. Even the weirdest arrangements and rules are usually there for a reason that (however inadequate) appeals to some people. Probably some people would be harmed by the discovery of a cure for cancer, though few would go so far as to complain.

Much of our welfare economics is built on thought experiments that involve paying compensation so that certain changes can be made to improve everyone's welfare. For instance, we argue that monopoly is Pareto-inefficient because in principle we could compensate the monopolist for a change to marginal cost pricing. But compensation is seldom paid when we do make changes. As a result, desirable changes are bitterly and often successfully opposed. Perhaps we need to think harder about when and how we should actually pay compensation.

References

Samuelson, Paul A., "The Consumer Does Benefit From Feasible Price Stability," *Quarterly Journal of Economics*, 1972, 86, 476–493.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 183–186

Recommendations for Further Reading

Bernard Saffran

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, the articles chosen will be expository or integrative and not focus on original research. As a rule of thumb, articles listed in this section appear in journals that are not indexed by the *Journal of Economic Literature* (which focuses on the research journals) and are similar to many of the articles published in this journal. They may include survey articles, discussion of related subjects in which economists might have an interest (for example, the use of game theory in evolutionary biology) or analyses of economics from other perspectives.

The intention is to publish a selective list of ten to fifteen articles per issue, with the selections depending heavily on input from readers. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Bernard Saffran, c/o *Journal of Economic Perspectives*, Department of Economics, Swarthmore College, Swarthmore, Pennsylvania 19081.

Hors d'oeuvres from the Federal Reserves

As always, the publications of the Federal Reserve banks offer some insightful examinations of current issues.

Loveman, Gary W., and Chris Tilly, "Good Jobs or Bad Jobs: What Does the Evidence Say?" *New England Economic Review*, January/February, 1988, 46–65. A useful survey that explores the view that the growth of employment has just led to new jobs that are bad jobs; if this view is correct, what causes this deterioration and what are the policy implications? They explore hypotheses that focus on structural and

■ *Bernard Saffran is Professor of Economics, Swarthmore College, Swarthmore, Pennsylvania.*

cyclical effects on labor supply and demand, as well as the argument that there has been no deterioration. They conclude, "There is thus general agreement on a decline in wage growth, an increase in the proportion of low-earnings jobs among full-time, year-round workers, and increased inequality within and between industries and occupations. However, the causes are not well understood... Consequently, the empirical evidence to date does not uniquely support any one policy agenda."

Meyer, Stephen A., "Non-Open-Market Monetary Policy Operations," *Business Review*, Federal Reserve Bank of Philadelphia, January/February 1988, 3–15. A discussion of the ways that the Fed influences the reserves of the banking system that do not rely on open-market operations: letting bills "run off," security transactions with foreign banks, and changes in the Treasury's balance. Useful in a "Money and Banking" course for its institutional content as well as providing for additional illustrations of the workings of the Fed's balance sheet.

Some Instabilities in Econometrics and Mathematics

A number of discussions have appeared on the relation of social science, mathematics and statistics.

Freedman, D. A., "As Others See Us: A Case Study in Path Analysis," *Journal of Educational Statistics*, Summer 1987. Freedman, who is Professor of Statistics, University of California, Berkeley, is a knowledgeable critic of econometric methodology. In this paper he criticizes the use of path analysis for analyzing social data. "At bottom, my critique is pretty simple-minded: Nobody pays much attention to the assumptions, and the technology seems to overwhelm common sense." There are comments from a number of social scientists and statisticians including Samuel Karlin, Thomas J. Rothenberg and Herman Wold. Rothenberg points out, for example, "Freedman wants to argue that the whole approach of structural model building is wrong-minded. He is mounting a full fledged attack on one of the dominant research methodologies in the social sciences. In particular, some of the most damaging missiles are aimed at my own field of econometrics." The entire issue is devoted to this discussion.

Koblitz, Neal, "A Tale of Three Equations; or The Emperors Have No Clothes;" Simon, Herbert A., "Unclad Emperors: A Case of Missing Identity," and a reply by Koblitz, *The Mathematical Intelligencer*, Volume 10, No. 1, 1988, 4–16. A continuation of the debate on the failure to admit Samuel Huntington, a Harvard political scientist, to the National Academy of Sciences. Koblitz is a mathematician at the University of Washington and Simon is a Nobel Laureate in Economics, winner of the National Medal of Science, a leading cognitive psychologist and, relevant in this context, the deliverer of the Josiah Willard Gibbs Lecture to the American Mathematical Society in 1984. Simon has written a defense of Huntington's mathematics in an unpublished article, "Some Trivial But Useful Mathematics," sections of which are republished here. Koblitz, who helped to initiate the attack on Huntington, now bares his fangs for Simon and there is a reply and rejoinder. Besides its obvious interest to those who enjoy academic blood sports or study the sociology of science, economists will find the technical discussion of ordinality and monotonic transformations useful.

Moore, David S., "Should Mathematicians Teach Statistics?" *The College Mathematics Journal*, January 1988, 3–35. Moore, Professor of Statistics at Purdue University, answers, "No! Statistics is no more a branch of mathematics than is economics, and should no more be taught by mathematicians... Its subject matter is data and inference from data. It is unprofessional for mathematicians who lack training and experience in working with data to teach statistics." Many of the issues raised relate to the teaching of economic statistics and indirectly to the teaching of economics. This article is then followed by a number of comments from mathematicians and statisticians, with a healthy sampling of those engaged in applied work.

Schools and Scholars

These articles range from Milton Friedman to James Meade to Walter Heller, and from public choice to institutional economics.

Breit, William, "Creating the 'Virginia School': Charlottesville as an Academic Environment in the 1960s," *Economic Inquiry*, October 1987, 645–657. An informal and anecdotal account of the personalities and politics in the development of the "public choice" and "constitutional economics" approaches at the University of Virginia.

Sargent, Thomas J., "Some of Milton Friedman's Scientific Contributions to Macroeconomics," Hoover Institution, Stanford University. Opening remarks to a symposium to celebrate Milton Friedman's seventy-fifth birthday. Sargent shows how Friedman's research on the consumption function, natural unemployment-rate hypothesis, frameworks and rules for macroeconomic stability, optimum quantity of money and monetary history have influenced and been modified by the later generation of economists. He concludes, "Many very good researchers continue to pay Milton Friedman the compliment of thinking hard about issues that he posed, often in terms that he defined, and using methods that he invented or inspired."

Solow, Robert M., "James Meade at Eighty," *The Economic Journal*, December 1986, 986–988. A warm tribute to James Meade, The Last Great Utilitarian who is concerned with Making Things Better: "If you have to be obsessed by something, maximizing real National Income is not a bad choice." I predict that the following paragraph will soon join the many other standard quotes by Solow: "When I emphasize this single-minded focus on economic policy, I do not refer to commentary on last month's doings in Whitehall. I suppose Meade's effort could be described as fundamental welfare economics, but welfare economics with red corpuscles, not the sort of attenuated theory that concludes that if only everything were convex and everybody knew everything and there were perfect markets for all future contingent commodities, including contingencies for which no vocabulary now exists, then with costless lump-sum transfers we could make all for the best in this best of imaginable worlds. Meade expects welfare economics to provide advice, not resignation."

Two longtime colleagues have written eulogies of Walter W. Heller. In *Challenge*, November–December 1987, pp. 59–64, James Tobin focuses on the innovative and effective Council of Economic Advisers that Heller led in Camelot and also includes a comparison of the economic performance of the Kennedy–Johnson and Reagan administrations during their first six years. Joseph A. Pechman, whose professional relationship with Heller dates from their graduate days at Wisconsin, writes in the *Brookings Papers on Economic Activity*, 1987, 2, pp. viii–xi. Pechman comments, "Walter Heller was one of the most important and effective economists of the post-World War II period. Yet he was not the greatest theorist, or the best teacher, or the most imaginative public official who ever held office. One is tempted to ask: what made him so effective? I think that the answer to that question is that he was so much greater than the sum of his separate skills. He joined high competence in all the dimensions of leadership with compassion, good judgement, and hard work, to leave a more indelible mark on the place of economics in American life than any other economist of his generation."

Two issues of the *Journal of Economic Issues* are devoted to institutional and evolutionary economics. The September 1987 issue deals with "Foundations of Institutional Thought," while the December 1987 issue deals with "Institutional Theory and Policy." The introduction explains, "These two volumes are intended to provide a comprehensive, contemporary formulation of institutional political economy as a viable alternative perspective to the prevailing neoclassical orthodoxy."

Calling on The Margin

The Margin. An interesting magazine aimed at an undergraduate economics audience. The December 1987 issue has a number of articles including a discussion of income distribution in China, elasticity, airline deregulation and an interview with Irma Adelman. There are also regular features on graduate schools, famous economists and even an economics crossword puzzle. Edited by Timothy Tregarthen. For a free sample copy, write to The Margin, Post Office Box 7150, Colorado Springs, Colorado, 80933-7150 or phone 1-800-Margin.

■ *I would like to thank Albert K. Smiley and Howard M. Wachtel for their suggestions.*

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 187–197

Anomalies
Cooperation

Robyn M. Dawes and Richard H. Thaler

Economics can be distinguished from other social sciences by the belief that most (all?) behavior can be explained by assuming that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear. An empirical result qualifies as an anomaly if it is difficult to "rationalize," or if implausible assumptions are necessary to explain it within the paradigm. This column will present a series of such anomalies. Readers are invited to suggest topics for future columns by sending a note with some references to (or better yet copies of) the relevant research. Comments on anomalies printed here are also welcome. The address is: Richard Thaler, c/o *Journal of Economic Perspectives*, Johnson Graduate School of Management, Malott Hall, Cornell University, Ithaca, NY 14853.

Introduction

Much economic analysis—and virtually all game theory—starts with the assumption that people are both rational and selfish. For example, predictions that players will defect in the prisoner's dilemma game and free ride in public goods environments are based on both assumptions. People are assumed to be clever enough to figure out that defection or free riding is the dominant strategy, and are assumed to care nothing for outcomes to other players; moreover, people are assumed to have no qualms about their failure to do "the right thing." [1]

---

[1] For a modern treatment of the theory of public goods, see Bergstrom, Blume, and Varian (1986).

■ *Robyn M. Dawes is Professor of Psychology and Head, Department of Social and Decision Sciences, Carnegie-Mellon University, Pittsburgh, PA. Richard H. Thaler is Henrietta Johnson Louis Professor of Economics at the Johnson Graduate School of Management, Cornell University, Ithaca, New York.*

The predictions derived from this assumption of rational selfishness are, however, violated in many familiar contexts. Public television successfully raises enough money from viewers to continue to broadcast. The United Way and other charities receive contributions from many if not most citizens. Even when dining at a restaurant away from home in a place never likely to be visited again, most patrons tip the server. And people vote in presidential elections where the chance that a single vote will alter the outcome is vanishingly small. As summarized by Jack Hirshleifer (1985, p. 55), "... the analytically uncomfortable (though humanly gratifying) fact remains: from the most primitive to the most advanced societies, a higher degree of cooperation takes place than can be explained as a merely pragmatic strategy for egoistic man." But why?

In this column and the next one, the evidence from laboratory experiments is examined to see what has been learned about when and why humans cooperate. This column considers the particularly important case of cooperation vs. free riding in the context of public good provision. (The next column is about the "ultimatum game.")

Single Trial Public Goods Experiments

To investigate why people cooperate, it is necessary to examine behavior in both single play and multiple play environments. Does cooperation evolve, for instance, only as individuals repeatedly interacting with each other find it in their interests to cooperate? A typical public goods experiment uses the following procedures. A group of subjects (most often students but sometimes other adult members of the community) is brought to the laboratory. Groups vary in size, but experiments usually have between 4 and 10 subjects. Each subject is given a sum of money, for example, $5. The money can either be kept and taken home, or some or all of the money can be invested in a public good, often called a "group exchange." Money invested in the group exchange for the n participants is multiplied by some factor k, where k is greater than 1.0 but less than n. The money invested, with its returns, is distributed equally among all group members. Thus, while the entire group's monetary resources are increased by each contribution (because $k > 1$), each individual's share of one such contribution is less than the amount she or he invests (because $k < n$). Suppose $k = 2$ and $n = 4$. Then if everyone contributes all $5 to the public good, each ends up with $10. This is the unique Pareto efficient allocation: no other solution can make everyone better off. But the dominant Nash strategy is to contribute nothing, because in exchange for a player's $5 contribution, that player receives only $2.50, while the rest of the payoff ($7.50) goes to the other players. The rational selfish strategy is to contribute nothing and hope that the other players decide to invest their money in the group exchange. If one player contributes nothing while all the others contribute $5, then that player will end up with $12.50, while the other players end up with $7.50. These conditions constitute a true social dilemma played with real money.

What does economic theory predict will happen in such a game? One prediction, called the *strong free rider hypothesis*, is that everyone will choose the dominant strategy, that is, nothing will be contributed to the public good. This is surely the outcome predicted by the selfish rational model. A less extreme prediction, called the *weak free rider hypothesis*, is that some people will free ride while others will not, yielding a suboptimal level of the public good, though not necessarily zero. The weak free rider hypothesis obviously does not yield very precise predictions.

The results of single play ("one shot") public goods experiments lend little support to the strong free rider hypothesis. While not everyone contributes, there is a substantial number of contributors, and the public good is typically provided at 40–60 percent of the optimal quantity. That is, on average, the subjects contribute 40–60 percent of their stake to the public good. In a study by Marwell and Ames (1981), these results held in many conditions: for subjects playing the game for the first time, or after a previous experience; for subjects who believed they were playing in groups of 4 or 80; and for subjects playing for a range of monetary stakes.[2] Indeed, Marwell and Ames found only one notable exception to this 40–60 percent contribution rate. When the subjects were a group of University of Wisconsin economics graduate students, the contribution rate fell to 20 percent, leading them to title their article "Economists Free Ride: Does Anyone Else?" [3] (Interestingly, economists told about the experiments predicted on the average a rate of about 20 percent—but for all participants, not just their students.)

Multiple Trial Experiments

A natural question to ask about the surprisingly high level of cooperation observed by Marwell and Ames is what would happen if the same players repeated the game several times. This question has been investigated by Kim and Walker (1984), Isaac, Walker, and Thomas (1984), and Isaac, McCue, and Plott (1985). The experimental design in these papers is similar to Marwell and Ames, except that there are usually 10 repetitions of the game. Two major conclusions emerge from these papers. First, on the initial trial, cooperation is observed at rates similar to those obtained by Marwell and Ames. For example, across 9 different experiments with varying designs, Isaac, McCue and Plott obtained a 53 percent contribution rate to the public good. Second, within a few repetitions, cooperation declines sharply. After 5 trials, the contributions to the public good were only 16 percent of the optimum. The experiments by Isaac, Walker and Thomas also obtained a decline in the contribution rate over time, though the decline was not as abrupt.[4]

---

[2] In the experiments with the highest stakes, contribution rates were somewhat lower, in the 28–35 percent range.

[3] This result has never been replicated, and so should be treated as preliminary. We wonder, however, whether economists *are* different. Do economists as a group donate less to charity than other similar groups? Are they less likely to leave tips in out-of-town restaurants?

[4] For experiments with a high return to contributing to the public good, the initial contribution rate was 52 percent, which fell to 32 percent on trial 10. In versions with low returns to contributing, the initial rate was 40 percent and the final rate was 8 percent.

Why does the contribution rate decline with repetition? One reasonable conjecture is that subjects learn something during the experiment that induces them to adopt the dominant strategy of free riding. Perhaps the subjects did not understand the game in the first trial and only learned that free riding was dominant over time. This possibility, however, appears unlikely in light of other experimental evidence. For example, the usual cooperation rates of roughly 50 percent are observed on trial one even for *experienced* subjects, that is, subjects who have participated in other multiple trial public goods experiments (e.g., Isaac and Walker, forthcoming). Also, Andreoni (1987a) has investigated the learning hypothesis directly, using the simple procedure of restarting the experiment. Subjects were told they would play a ten-period public goods game. When the ten periods were completed, the subjects were told they would play again for another ten rounds with the same other players. In the first ten trials Andreoni replicated the decaying contribution rate found by previous investigators, but upon restarting the game contributions went back up to virtually the same contribution rates observed on the initial trial in the first game (44 percent on trial one of the second game vs. 48 percent in the first). Such results seem to rule out any explanation of cooperation based on subjects' misunderstanding the task.[5]

Reciprocal Altruism

One currently popular explanation of why we observe so much cooperation in and outside of the laboratory invokes reciprocal altruism as the mechanism. This explanation, most explicitly developed by Axelrod (1984), is based on the observation that people tend to reciprocate—kindness with kindness, cooperation with cooperation, hostility with hostility and defection with defection. Thus, being a free rider may actually be a less fruitful strategy when the chooser takes account of the probable future response of others to his or her cooperation or defection. A cooperative act itself —or a reputation for being a cooperative person—may with high probability be reciprocated with cooperation, to the ultimate benefit of the cooperator.

The most systematic strategy based on the principle of reciprocal altruism is a TIT-FOR-TAT one first suggested by Anatol Rapoport, in which a player begins by cooperating and then chooses on trial t the same response the other player has made on trial $t - 1$. The real strength of this explanation lies in demonstrating, both analytically and by computer tournaments of interacting players (programs) in iterated social dilemmas, that any person or small group of people practicing such reciprocal altruism will have a statistical tendency to receive higher payoffs "in the long run" than those who don't practice it. In fact, TIT-FOR-TAT "won" two computer tournaments Axelrod conducted in which game theorists proposed various strategies that were compared against each other in pairwise encounters with repeated plays. Because evolution is concerned with such long run probabilistic phenomena, it can be inferred that reciprocating people have greater "inclusive fitness" than dc non-reciprocating ones. Hence, to the degree to which such a tendency has some genetic basis, it should evolve as an adaptation to the social world.

---

[5] A similar conclusion is reached by Goetze and Orbell (forthcoming).

An implication of reciprocal altruism is that individuals will be uncooperative in dilemma situations when there is no possibility of future reciprocity from others, as in situations of anonymity or interacting with people on a "one-shot" basis. Yet we observe 50 percent cooperation rates even in single trial experiments, so reciprocal altruism cannot be used directly to explain the experimental results described so far. Also, of course, it is very difficult to play TIT-FOR-TAT, or any other strategy based on reciprocal altruism, when more than 2 people are involved in the repeated dilemma situation. If some members of a group cooperate on trial t while others defect, what should a player attempting to implement a TIT-FOR-TAT type strategy choose on the subsequent trial?

A related hypothesis that appears consistent with the decaying contribution rates observed in the multiple trial experiments is suggested by the theoretical work of Kreps, Milgrom, Roberts, and Wilson (1982). They investigate the optimal strategy in a repeated prisoner's dilemma game with a finite number of trials. If both players are rational, then the dominant strategy for both is to defect on every trial. While TIT-FOR-TAT has been shown to be effective in infinitely repeated prisoner's dilemma games (or equivalently, games with a constant small probability of ending after any given trial), games with a known end point are different. In any finite game both players know that they should defect on the last trial, so there is no point in cooperating on the penultimate trial, and by backward induction, it is never in one's best interest to cooperate. What Kreps et al. show is that if you are playing against an opponent whom you think may be irrational (i.e., might play TIT-FOR-TAT even in a game with finite trials), then it may be rational to cooperate early in the game (to induce your irrational opponent to cooperate too). Since the public goods games have a similar structure, it could be argued that players are behaving rationally in the Kreps et al. sense. Once again, however, the data rule out this explanation. Cooperation never falls to zero, even in one-trial games or in the last period of multi-trial games when it can never be selfishly rational to cooperate.

Additional evidence against the reciprocity hypothesis comes from another experiment designed by Andreoni. One group of 15 subjects played repeated trials in groups of 5 as described above. Another group of 20 subjects played the same game in groups of 5, but the composition of the group varied on each trial. Moreover, the subjects did not know which 4 of the other 19 subjects would constitute their group in any given round of the game. In this condition, there can be no strategic advantage to cooperation, since the players in the next round will be, in essence, strangers. If cooperation in early rounds of these experiments is observed, strategic cooperation can be ruled out. Indeed, Andreoni found that cooperation was actually a bit higher in the stranger condition than in a comparable condition where the groups remained intact. (This effect was statistically significant, though slight.)

One conclusion which emerges from these experiments is that people have a tendency to cooperate until experience shows that those with whom they are interacting are taking advantage of them. This "norm of cooperation" will resemble reciprocal altruism in infinitely repeated games; but the behavior, as we have seen, is also observed in cases when reciprocal altruism would be inappropriate. One explanation for this type of behavior is offered by Robert Frank (1987). Frank argues that people who adopt a norm of cooperation will do well by eliciting cooperation from others, and attracting interaction with other cooperators. The key to Frank's argument is that one cannot successfully fake being cooperative for an extended period of time—just as one cannot be successful getting people to believe too many lies.[6] Furthermore, because cooperators are, by assumption, able to identify one another, they are able to interact selectively and exclude defectors.

Altruism

There are other explanations of why people cooperate both in the lab and the field. One is that people are motivated by "taking pleasure in others' pleasure." Termed *pure altruism*,[7] this motive has been eloquently stated by Adam Smith, in *The Theory of Moral Sentiments* (1759; 1976): "how selfish soever man may be supposed to be, there are evidently some principles in his nature, which interest him in the fate of others, and render their happiness necessary to him, though he derive nothing from it, except the pleasure of seeing it." While the pleasure involved in seeing it may be considered "selfish" (following the sophomoric argument that altruism is by definition impossible, because people do what they "want" to do), the passage captures the idea that people are motivated by positive payoffs for others as well as for themselves. Consequently, they may be motivated to produce such results through a cooperative act. One problem with postulating such pure altruism as a reason for contributing to public goods is that such contributions cannot be explained purely in terms of their effects. If they could, for example, then governmental contributions to the same goal should "crowd out" private contributions on a dollar-for-dollar basis, since the results are identical no matter where the funding comes from. Such crowding out does not appear to be nearly complete. In fact, econometric studies indicate that an increase in governmental contribution to such activities is associated with a decrease in private contribution of only 5 to 28 percent (Abrams and Schmitz, 1978, 1984; Clotfelter, 1985).

Another type of altruism that has been postulated to explain cooperation is that involved in the act of cooperating itself, as opposed to its results. "Doing the right (good, honorable,...) thing" is clearly a motive for many people. Sometimes termed *impure altruism*, it generally is described as satisfaction of conscience, or of noninstrumental ethical mandates.

The roles of pure and impure altruism and other causes for cooperation (or the lack thereof) have been investigated over the last decade by the team of Robyn

---

[6] As the late Senator Sam Ervin said: "The problem with lying is that you have to have a perfect memory for what you said." None of us do. It's easier to remember what actually happened, although that is not easy either.

[7] The terms pure and impure altruism are introduced by Andreoni (1987b).

Dawes, John Orbell and Alphons van de Kragt. In one set of experiments (Dawes et al., 1986), they examined the motives for free riding. The game used for these experiments had the following rules. Seven strangers were given $5 each. If enough people contributed their stake to the public good (either 3 or 5 depending on the experiment), then every person in the group would receive a $10 bonus whether or not they contributed. Thus, if enough subjects contribute, each contributor would leave with $10 and each non-contributor would leave with $15. If too few contributed, then non-contributors would keep their $5 while contributors would leave with nothing. Subjects were not permitted to talk to one another (though this was modified in subsequent experiments). In this context two reasons for not contributing can be identified. First, subjects may be afraid that they will contribute but not enough others will, so their contribution will be futile. This motive for defecting was termed "fear." Second, subjects may hope that enough others will contribute and hope to receive $15 instead of $10. This motive was called "greed." The relative importance of fear and greed was examined by manipulating the rules of the game. In the "no greed" condition, payoffs were changed so that all subjects would receive $10 if the number of contributors was sufficient (rather than $10 for contributors and $15 for free riders). In the "no fear" condition contributors were given a "money back guarantee:" if a subject contributed and not enough others did, the subject would receive the money back. (However, in this condition if the public good was provided, contributors would receive only $10 while free riders would get $15.) The results suggested that greed was more important than fear in causing free riding. In the standard game contribution rates averaged 51 percent. In the no fear (money back) game contributions rose to 58 percent, but in the no greed game contributions were 87 percent.[8]

Another possible interpretation is that the no greed condition can produce a stable equilibrium, while the no fear cannot. If subjects in the no greed condition believe that the mechanism of truncating payoffs works to motivate others to contribute, their motive will be enhanced as well, because the only negative result of contributing occurs if enough others *don't* contribute. In contrast, subjects in the no fear condition who conclude that the conditions will encourage others to contribute will be tempted to free ride themselves, leading to the conclusion that others will be tempted as well, leading to the conclusion that they should themselves contribute, etc.—an infinite loop.

One of the most powerful methods for inducing cooperation in these games is to permit the subjects to talk to one another. Twelve groups were run with the same payoffs described earlier, but under conditions in which discussions were allowed. The effect of this discussion was remarkable (van de Kragt, et al., 1983). Every group used the discussion period to specify a group of people who were designated to cooperate. The most common means of making the distributional decision was by lottery, though

[8] Notice that contributing could be selfishly rational if a subject thought that the probability his or her contribution would be critical (that is, exactly $M - 1$ others will contribute) was greater than one-half. However, subjects who contributed did not generally believe that their contribution was necessary. Virtually no contributors believed they were critical to obtaining the public good with a probability greater than 0.50. In fact, pooling across all conditions, 67 percent of the contributors believed so many others would contribute that their own contributions would be redundant.

volunteering was also observed. One group attempted interpersonal utility comparisons to determine relative "need." Whatever methods the groups used, they worked. All 12 groups provided the public good, and in 3 of the groups more than the required number of subjects contributed. These results are consistent with the earlier ones. Subjects designated as contributors cannot greedily expect more from free riding, because their contributions are (believed to be) crucial for their obtaining the bonus (and were in all but 3 groups). Moreover, belief that others in the designated set of contributors will be motivated to contribute by the designated contributor mechanism will enhance—rather than diminish—each designated contributor's motive to contribute.

One possible explanation for the value of discussion is that it "triggers" ethical concerns that yield a utility for doing the "right" thing (that is, impure altruism). Elster (1986), for example, has argued that group discussions in such situations yield arguments for group-regarding behavior (it is hard to argue for selfishness), and that such arguments have an effect not only on the listener but on the person making them as well. To test this hypothesis, a new set of experiments was conducted (Orbell, van de Kragt and Dawes, forthcoming). In this set of experiments all 7 subjects were given $6 each. They could either keep the money or contribute it to the public good in which case it would be worth $12 to the other 6 members of the group. In this case, keeping the $6 is a dominant strategy because the person who does so receives both that $6 and $2 from each of the other 6 group members who gave away the money.

Subjects first met in groups of 14 in a waiting room in which they were not allowed to talk; they were then divided into the two groups on a clearly random basis. Half of these subgroups were allowed to talk about the decision, half not. The experimenters told half of the groups that the $12 given away would go to the other six people in their own group, while the other half of the groups were told that the money would go to six people in the other group. There are thus four conditions—discussion or no discussion crossed with money goes to own group or other group. If discussion simply makes individuals' egoistic payoffs clear, then it should not increase cooperation rate in any of these conditions since free riding is dominant. If, however, discussion increases utility for the act of cooperation per se, then discussion should be equally effective whether the money given away goes to members of their own group or to the other group—which consists, after all, of very similar people who were indistinguishable prior to the random drawing (usually college students or poorer members of the community).

The results were clear. In the absence of discussion, only about 30 percent of the subjects gave away the money, and those who did so indicated that their motive was to "do the right thing" irrespective of the financial payoffs.[9] Discussion raises the cooperation rate to 70 percent, but only when the subjects believe the money is going to members of their own groups; otherwise, it is usually less than 30 percent. Indeed, in such groups it was common to hear comments that the "best" possible outcome would be for all group members to keep their money while those in the other group gave it away (again, people from whom the subjects have been randomly separated about 10 minutes earlier).

---

[9] In a similar—but simulated—one-shot experiment Hofstadter (1983) had discovered a roughly identical cooperation rate among his eminent friends. Most defect, but some cooperate, and for reasons of impure altruism. As one cooperator, Professor Daniel C. Dennett of Tufts, put it: "I would rather be the person who bought the Brooklyn bridge than the person who sold it. Similarly, I feel better spending $3 gained by cooperation than $10 gained by defection." (Hofstadter terms that a "wrong reason" for cooperating in a dilemma situation; yet it is the one often given by the subjects who cooperate without discussion in the experiments described above, and similar ones.)

Thus, group identity appears to be a crucial factor in eschewing the dominating strategy. That result is compatible with previous social-psychological research on the "minimal group" paradigm (e.g. Tajfel and Turner, 1979; the papers contained in Turner and Giles, 1981), which has repeatedly demonstrated that allocative decisions can be sharply altered by manipulations substantially weaker than 10 minutes of discussion. For example, a "common fate" group identity—where groups received differing levels of payoffs depending on a coin toss—led subjects to attempt to "compensate" for non-cooperators in their own group by increasing cooperation rates, while simultaneously decreasing cooperation when the non-cooperators were believed to be in the other group, even when the identities of the people involved were unknown (Kramer and Brewer, 1986).

In the groups in which discussion was permitted it was very common for people to make promises to contribute. In a second series of experiments, Orbell, van de Kragt and Dawes investigated whether these promises were important in generating cooperation. Perhaps people feel bound by their promises—or believe they will receive a "satisfactory" payoff if they give away the money when others promise to do so because others will be bound by such promises. The main result was that promise making was related to cooperation only when every member of the group promised to cooperate. In such groups with universal promising, the rate of cooperation was substantially higher than in other groups. In groups in which promising was not universal there was no relationship between each subject's choice to cooperate or defect and (1) whether or not a subject made a promise to cooperate, or (2) the number of other people who promised to cooperate. Consequently, the number of promises made in the entire group and the group cooperation rate were unrelated. These data are consistent with the importance of group identity if (as seems reasonable) universal promising creates—or reflects—group identity.

Commentary

In the rural areas around Ithaca it is common for farmers to put some fresh produce on a table by the road. There is a cash box on the table, and customers are expected to put money in the box in return for the vegetables they take. The box has just a small slit, so money can only be put in, not taken out. Also, the box is attached to the table, so no one can (easily) make off with the money. We think that the farmers who use this system have just about the right model of human nature. They feel that enough people will volunteer to pay for the fresh corn to make it worthwhile to put it out there. The farmers also know that if it were easy enough to take the money, someone would do so.

In contrast to these farmers, economists either avoid judgments of human nature, or make assumptions that appear excessively harsh. It is certainly true that there is a "free rider problem." Not all people can be expected to contribute voluntarily to a good cause, and any voluntary system is likely to produce too little of the public good (or too much of the public bad in the case of externalities). On the other hand, the strong free rider prediction is clearly wrong—not everyone free rides all of the time.

There is a big territory between universal free riding and universal contributing at the optimal rate. To understand the problems presented by public goods and other dilemmas it is important to begin to explore some issues that are normally ignored in economics. For example, what factors determine the rate of cooperation? It is encouraging to note that cooperation is positively related to the investment return on the public good. The more the group has to gain through cooperation, the more cooperation is observed—the supply of cooperation is upward sloping. The results involving the role of discussion and the establishment of group identity are, however, more difficult to incorporate into traditional economic analyses. (One economist attempting to do so proposed that group discussion simply confuses subjects to the point that they no longer understand it is in their best interests to be defectors.)

More generally, the role of selfish rationality in economic models needs careful scrutiny. Amartya Sen (1977) has described people who are always selfishly rational as "rational fools," because mutual choices based only on egoistic payoffs consistently lead to suboptimal outcomes for all involved. Perhaps we need to give more attention to "sensible cooperators."

■ *We wish to thank James Andreoni, Linnda Caporael, Mark Isaac, and John Orbell for helpful comments on an earlier draft.*

References

Abrams, Burtran A., and Mark A. Schmitz, "The Crowding Out Effect of Government Transfers on Private Charitable Contributions," *Public Choice*, 1978, *33*, 29–39.

Abrams, Burtran A., and Mark A. Schmitz, "The Crowding Out Effect of Government Transfers on Private Charitable Contributions: Cross Sectional Evidence," *National Tax Journal*, 1984, *37*, 563–68.

Andreoni, James, "Why Free Ride? Strategies and Learning in Public Goods Experiments," unpublished, University of Wisconsin, Department of Economics, 1987a.

Andreoni, James, "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," unpublished, University of Wisconsin, Department of Economics, 1987b.

Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.

Bergstrom, Theodore, Lawrence E. Blume, and Hal Varian, "On the Private Provision of Public Goods," *Journal of Public Economics*, 1986, *29*, 25–49.

Clotfelter, Charles T., *Federal Tax Policy and Charitable Giving*, Chicago: The University of Chicago Press, 1985.

Dawes, Robyn M., John M. Orbell, Randy T. Simmons, and Alphons J. C. van de Kragt, "Organizing Groups for Collective Action," *American Political Science Review*, 1986, *80*, 1171–1185.

Elster, Jon, "The Market and the Forum: Three Varieties of Political Theory." In Jon Elster and Aanund Hylland, eds., *Foundations of Social Choice Theory: Studies in Rationality and Social Change*. Cambridge: Cambridge University Press, 1986, 103–132.

Frank, Robert, "If *Homo Economicus* Could Choose his own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, September 1987, *77*, 593–605.

Goetze, David, and John M. Orbell, "Understanding and Cooperation," *Public Choice*, forthcoming.

Hirshleifer, Jack, "The Expanding Domain of Economics," *American Economic Review*, December 1985, *75*, *Number 6*, 53–70.

Hofstadter, Douglas, "Metamagical Themas," *Scientific American*, 1983, *248*, 14–28.

Isaac, R. Mark, Kenneth F. McCue, and Charles Plott, "Public Goods Provision in an Experimental Environment," *Journal of Public Economics*, 1985, *26*, 51–74.

Isaac, R. Mark, James M. Walker, and Susan H. Thomas, "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations," *Public Choice*, 1984, *43*, 113–149.

Isaac, R. Mark, and James M. Walker, "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics*, forthcoming.

Kim, Oliver, and Mark Walker, "The Free Rider Problem: Experimental Evidence," *Public Choice*, 1984, *43*, 3–24.

Kramer, R. M., and Marilyn Brewer, "Social Group Identity and the Emergence of Cooperation in Resource Conservation Dilemmas." In H. Wilke, D. Messick, and C. Rutte, eds., *Psychology of Decision and Conflict. Vol.* 3. *Experimental Social Dilemmas*, Frankfurt Am Main: Verlag Peter Lang, 1986, pp. 205–230.

Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson, "Rational Cooperation in Finitely Repeated Prisoners' Dilemmas," *Journal of Economic Theory*, 1982, *27*, 245–252.

Marwell, Gerald and Ruth Ames, "Economists Free Ride, Does Anyone Else?" *Journal of Public Economics*, 1981, *15*, 295–310.

Orbell, John M., Robyn M. Dawes, and Alphons J. C. van de Kragt, "Explaining Discussion Induced Cooperation," *Journal of Personality and Social Psychology*, forthcoming.

Sen, Amartya K., "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Journal of Philosophy and Public Affairs*, 1977, *6*, 317–344.

Smith, Adam, *The Theory of Moral Sentiments*. Oxford: Clarendon Press, 1976. (Originally published in 1759.)

Tajfel, Henri and John C. Turner, "An Integrative Theory of Intergroup Conflict." In W. Austin and S. Worchel, eds., *The Social Psychology of Intergroup Relations*, Monterey, CA: Brooks/Cole, 1979, pp. 33–47.

Turner, John C., and Howard Giles, *Intergroup Behavior*, Chicago: University of Chicago Press, 1981.

van de Kragt, Alphons J. C., John M. Orbell, and Robyn M. Dawes, "The Minimal Contributing Set as a Solution to Public Goods Problems," *American Political Science Review*, 1983, *77*, 112–22.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 199–203

Correspondence

To be considered for publication in the Correspondence section, letters should be relatively short—generally less than 1000 *words—and should be sent to the journal offices at the address appearing inside the front cover. The editors will choose which letters will be published. All published letters will be subject to editing for style and length.*

Comments on "Symposium on Mergers and Antitrust," Fall 1987, pp. 3–54

This journal's recent symposium on "Horizontal Mergers and Antitrust" leads the reader to a rather surprising conclusion: that none of the participating experts is willing and able to recommend a merger policy which he can consistently defend on its economic merits. Thus, all three authors recognize that beneficial results of mergers (such as cost reductions, or efficiencies) are as relevant to their total economic effect as are detrimental results (such as tacit collusion). At the same time, all three recommend policies under which the *legality* of a merger is entirely, or in most cases, unaffected by its expected beneficial consequences.

Professor White, for example, recommends that efficiencies be ignored in practice, largely because he believes that available evidence is likely to give an inaccurate impression of the ultimate results and also, secondarily, because ignoring efficiencies seems "to be consistent with the Supreme Court's interpretation of the Clayton Act" (p. 18). Obviously, neither of these reasons casts any doubt on the economic relevance of efficiencies.

Professor Schmalensee believes that enforcement agencies deciding whether or not to oppose a merger should consider evidence on efficiencies, but only as "tiebreakers"; that is, that "strong efficiency arguments should tip the scales in close cases" (p. 44). He rejects any broader use of efficiencies data by these agencies because of both inadequate information and the shortcomings of economic analysis. The courts, he believes, should be barred from considering such data because they cannot cope with "complexity and tradeoffs" (p. 42).

Professor Fisher would give merger proponents "the burden of proving that the merger will create efficiencies to offset any increased risk of anticompetitive behavior," with a "very high standard" of proof required because "[s]uch claims are easily made and all too easily believed" (pp. 26–27, 36). At the end of the analysis, however, it becomes clear that, in Fisher's view, prospective efficiencies should play little or no part in determining legality.

Having left the efficiencies question in unjustified limbo, none of the authors can argue that his policy is designed to produce net economic benefit. Largely because they are difficult to estimate, the benefits of mergers are not allowed to affect antitrust policy. On the other hand, the much more serious difficulties involved in estimating losses from expected increases in market power do not deter the authors from recommending action to prevent these losses, regardless of possible merger benefits.

Clearly, difficulties of measurement do preclude calculation of the net economic benefits to be expected from proposed mergers; antitrust policies relying on this type of calculation must be ruled out. However, there may be other ways of approaching the problem—ways that are workable as well as defensible on economic grounds. For example, empirical research may identify classes of anticompetitive merger with practically negligible benefits, in the sense that these benefits have not proven to be large enough to offset the costs of screening individual proposals to reveal them. Here there would be no need for *ex ante* estimation of benefits of individual merger proposals and, of course, no need for any estimates of losses from reduction of competition. These classes could be banned. A similar criterion of legality could (at least in principle) be applied on a case-by-case basis: any anticompetitive merger not shown likely to produce overall efficiencies (cost reductions exceeding cost increases) could be simply rejected. Only rough estimates of benefits and no estimates of competitive losses would be involved. Both in identifying appropriate classes and in evaluating individual proposals, efficiencies obtainable by means less restrictive than merger would of course be disregarded, since rejection of the merger alternative would not cause loss of the efficiencies.

This policy would not seek to ensure that every—or indeed, any—permitted merger would engender benefits outweighing losses due to reduced competition. Compared to actual judgments based on perfect foresight, the policy could produce only worse results. Its more modest aim is to prevent anticompetitive mergers with neutral or negative effects on efficiency. Nothing is lost by banning them and there may be significant gains. Compared to simple inaction, therefore, the policy can be defended on economic grounds.

It also appears to be the only course consistent with the law's present treatment of enterprise expansion by internal growth. Such growth is not prevented whenever it eliminates competition or results in a certain degree of market concentration (with a certain degree of entry difficulty and a certain mix of factors tending to foster collusion). Nor is such growth subjected to a requirement that efficiencies outweigh anticompetitive effects. On the contrary, growth is freely permitted (at least in the absence of monopolistic intent or exclusionary practices) presumably on the sound premise that it can usually occur only in connection with superior economic performance—that is, by attracting more purchasers through offering lower prices or more satisfactory products. The suggested rule for anticompetitive mergers is intended to ensure that they, too, produce some economic benefits.

Lucile S. Keyes
Washington, D.C.

When I first read that the AEA planned to put out a new journal that would communicate developments in economics to nonspecialists, I thought it was good idea. However, I have now seen two issues of the journal, and I have changed my mind.

I have several reactions to the symposium on horizontal mergers and antitrust. First, nothing said appears to be original or even state of the art. Second, Fisher's article was a personal attack and thus inappropriate. Finally, it is quite odd that no one from an enforcement agency was invited to contribute; indeed, I should have been invited to participate. The basic subject of the symposium was the Merger Guidelines and their appreciation, particularly with respect to market delineation. I would have been a logical person to participate because I invented, know more than any one else about, and published more than anyone else about, the analytical paradigm of the Merger Guidelines and Guidelines' approach to market delineation.

Had I been invited to participate, among other things, I would have commented on remarks by Fisher and Schmalensee about the Guidelines' use of the prevailing price, as opposed to the competitive price, in market delineation. Fisher and Schmalensee pose the hypothetical of tacit collusion that might break down if not for a proposed merger. They contend that the Merger Guidelines are likely to cause this anticompetitive merger to go unchallenged because of the Guidelines' use of the prevailing price. They also contend that using the competitive price instead solves the problem. My comments are that the antitrust enforcers could not know that a proposed merger fits this hypothetical, but if they did, the Guidelines would give exactly the answer Fisher and Schmalensee want to get.

Consider the problem faced by enforcers confronted with a proposed merger that they must analyze. They could not know that the firms in the industry are tacitly colluding, but let us assume that this information is revealed. Now, if the collusion would continue with or without the merger, the Guidelines get the right answer and the competitive price advocates get the wrong answer, so assume that the collusion would continue only with the merger. Of course, the enforcers could not know that the tacit collusion will continue only with the merger, but we can assume that thev do. Now after all this assuming, Fisher and Schmalensee (and many others I might add) would say that the Merger Guidelines get the wrong answer. However, they overlook the fact that the Guidelines prescribe the use of prevailing *or likely future prices* to handle this contingency (and a few others). This is explained in two of my articles on the Guidelines (1983, pp. 525–526; 1985, pp. 651–653). If Fisher and Schmalensee do not find my argument persuasive, their critique of it would have made their articles much more interesting.

Gregory J. Werden
Antitrust Division
U.S. Department of Justice

References

Fisher, Franklin M., "Horizontal Mergers: Triage and Treatment," *Journal of Economic Perspectives*, Fall 1987, *1*, 23–40.

Schmalensee, Richard, "Horizontal Merger Policy: Problems and Changes," *Journal of Economic Perspectives*, Fall 1987, *1*, 41–54.

Werden, Gregory J., "Market Delineation and the Justice Department's Merger Guidelines," *Duke Law Journal*, June 1983, 514–579.

Werden, Gregory J., "A Closer Analysis of Antitrust Markets," *Washington University Law Quarterly*, No. 4, 1985, *62*, 647–669.

Comment on Symposium on Arbitrage, Fall 1987, pp. 55–94.

It was with immense satisfaction that I read the objectives of the new *Journal*, to disseminate knowledge broadly and nontechnically, in its first issue. It was with overwhelming disappointment that I noted the gross violation of these objectives in the second issue, specifically 18 pages of arcanum by Hal Varian. Ironically, the survey by Klamer and Colander in the same issue reveals a corresponding sacrifice of knowledge for technique worship in graduate economics education.

Thomas Havrilesky
Professor of Economics
Duke University
Durham, North Carolina

Response from the editors: The *Journal of Economic Perspectives* was founded to serve a number of different functions. One of the explicit objectives of the new journal was to make more readily accessible important new findings and to draw attention to new directions of research, much as *Scientific American* does for the biological and physical sciences. Much of the original research upon which these papers will be based uses quite advanced mathematics. The authors are encouraged to make the articles as accessible as possible, since they are intended for the broad range of economists, but doing so is not an easy task.

Inevitably, some of the articles will require the kind of mathematics expected of current first year graduate students, and even then, the article will not make good bedside reading. But if the article has facilitated the dissemination of important ideas we feel it has accomplished its objectives. Mark Machina's article on behavior under uncertainty in the first issue, and the Varian-Rubinstein symposium on option pricing and derivative assets in the second represent our first attempts at this kind of exposition. We certainly do not plan to fill the journal with articles of this type, but we feel strongly that they have a place in the journal in limited quantities.

Professor Havrilesky suggests that Hal Varian's article on arbitrage is "arcanum." That misses the point in two ways. First, the very purpose of the article is precisely to attempt to make difficult topics *more* accessible, and we believe the Varian and Rubinstein articles accomplished that goal. Second, this "arcanum" is believed responsible by many people for the stock market crash of last October. Portfolio insurance and futures trading, which are inextricably tied to explaining the crash, are important examples of derivative assets. The Varian and Rubinstein articles not only explained what derivative assets (including portfolio insurance) were, but explained how they were constructed and priced.

We would like to assure Havrilesky and others that we are committed to doing everything we can to ensure that the articles we publish are as clearly and simply written as possible.

Addendum from Hal Varian: In my recent exposition of "The Arbitrage Principle in Financial Economics," I described Breeden and Litzenberger's (1978) article which showed how options can be used to construct Arrow-Debreu securities. However, I should have also mentioned Ross's (1976) article which contained an early demonstration of how options can be used to create securities that can span a state space. Dybvig and Ross (1987) contains a more advanced discussion of arbitrage techniques along with a more detailed bibliography, and is a good source for further reading in this area.

References

Breeden, Douglas, and Robert Litzenberger, "Prices of State Contingent Claims Implicit in Option Prices," *Journal of Business*, 1978, *51*, pp. 621–651.

Dybvig, Philip, and Stephen Ross, "Arbitrage," *The New Palgrave*, New York: The Stockton Press, 1987, 100–106.

Ross, Stephen, "Options and Efficiency," *Quarterly Journal of Economics*, 1976, *40*, 75–89.

Journal of Economic Perspectives—Volume 2, Number 3—Summer 1988—Pages 204–207

Notes

The annual meeting for the American Economic Association will be held in New York, NY, December 27–30, 1988.

The Professional Placement Service will be held at the New York Marriott Marquis Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30.

Members wishing to give papers or make suggestions for the program for the AEA meeting, Atlanta, GA, December 27–30, 1989, are invited to write to Elton Hinshaw, American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. Include the bibliographic code (*Journal of Economic Literature* codes) for each suggested topic, paper proposal and abstract. All proposals will be grouped by the *JEL* code and sent to the appropriate member of the Program Committee. To be considered for contributed sessions, abstracts of (non-econometric) papers must be received no later than February 1, 1989.

The *Journal of Economic Perspectives* has relocated its editorial offices. Its new address is: *The Journal of Economic Perspectives*, Department of Economics, Stanford University, Stanford, CA 94305-6072.

The American Economic Association has renewed its agreement with Hertz. Members are eligible for the following discounts on car rentals. In the United States, excluding Puerto Rico, they are: a) 10% on Affordable Daily Rates (all car classes); b) 10% on Affordable Weekly Rates (compact and larger cars); c) 5% on Affordable Weekend Rates (compact and larger cars); d) these discounts are available at all participating locations. In Canada, the discount is 15% on published "Basic Time and Kilometer" rates, except rentals of Class A cars. In International travel, the discount on published "Basic Time and Kilometer" rates, or "SUM Daily Rates" (if applicable), where such discounts are permitted by law, follows: 10% at all participating International Locations except in Eastern Europe and Iceland, where a 5% discount shall apply.

You will receive a new Hertz identification card in the registration packet for the 1988 annual meetings. Members in countries other than the United States, Canada and Mexico should write to the following address to request a card: Card, American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786.

The University of Maryland has an electronic bulletin board, Maryland EConomic CAll (MECCA), that provides thousands of the most used economic time series in a uniform format suitable for economic analysis. MECCA is operated by the University's Computer Laboratory of the College of Behavioral and Social Sciences in collaboration with the Inforum research project and the Department of Economics. Initially no service charge is made. The access number is 301-454-7326 (2400, 1200 or 300 baud, 8-bits, no parity, 1 stop bit). One needs a microcomputer with DOS and a modem connected to a telephone to have access to over 10,000 economic time series. MECCA may charge for access to banks which are costly. No access charge will be made for data given to MECCA by the producing agency. If usage grows and telephone service must be expanded, a small subscription or connect-time charge may be necessary. MECCA welcomes secondary distribution of its free material by other bulletin boards. For more information, contact Clopper Almon or Robert Bennett, Department of Economics, University of Maryland, College Park, MD 20742. Phone: 301-454-5384.

The **Council for International Exchange of Scholars** has announced the opening of competition for the 1989–90 Fulbright grants in research and university lecturing abroad. The awards for 1989-90 include more than 300 grants in research and 700 grants in university lecturing for periods from three months to a full academic year. Openings exist in over 100 countries. Applications are encouraged from retired faculty and independent scholars. Eligibility requirements for a Fulbright Award are U.S. citizenship, Ph.D. or comparable professional qualifications, university or college teaching experience, and, for selected assignments, proficiency in a foreign language. Application deadline for the Awards are: September 15, 1988, for Africa, Asia, Europe, the Middle East, and lecturing awards to Mexico, Venezuela, and the Caribbean; November 1, 1988, for institutional proposals for, the Scholar-in-Residence Program; January 1, 1989, for Administrators' Awards in Germany, the United Kingdom, and Japan; the Seminar in German Civilization; the NATO Research Fellowships, and the Spain Research Fellowships; and February 1, 1989, for the France, Italy and Germany Travel-Only Awards. For more information and applications, contact: Council for International Exchange of Scholars, Eleven Dupont Circle N. W., Washington, DC 20036-1257. Phone: 202-939-5403.

Nominations for candidates for **Sloan Research Fellowship** are due by September 15. Candidates must be members of the regular faculty at a college or university in the United States or Canada and must be at an early stage of their research careers. For information write: Sloan Research Fellowship, Alfred P. Sloan Foundation, Suite 2550, 630 Fifth Avenue, New York, NY 10111.

Call for papers. The **Southwestern Society of Economists** will meet in conjunction with the 16th Annual Meeting of the Southwestern Federation of Administrative Disciplines (SWFAD) in New Orleans, March 8–11, 1989. Proposals of 250 words in any area of economics are welcome. Include the proposal abstract and a cover letter with name, affiliation and position, address, telephone number, and area that you could serve as a discussant if your proposal is accepted for the program. Papers and discussant comments may be included in the *Journal of the Southwestern Society of Economists*. Prior to October 1, 1988, submit materials to Jim T. McMinn, SSE Program Chair, Department of Economics and Finance, Austin Peay State University, Clarksville, TN 37040.

Call for papers. The **Business Association of Latin American Studies** will hold a conference February 15–18, 1989, in Boca Raton, Florida, to discuss issues and research that affect the economies and business environment of Latin America and the Caribbean. Papers, panel proposals, and abstracts, transmitted in quadruplicate and subject to blind review, are being accepted until October 8, 1988. For more information, write: Dr. Robert P. Vichas, Florida Atlantic University, P.O. Drawer 7638, Fort Lauderdale, Florida 33338.

Call for papers. The **Midsouth Academy of Economics & Finance** will hold its 16th annual meeting in Nashville, TN, February 15–18, 1989. All topics in economics and finance are appropriate. Send two copies of a 500-word abstract or two copies of the completed paper by October 15, 1988, to the General Program Coordinator. Papers and comments presented by registrants may be published in the papers and proceedings section of *The Journal of Economics and Finance*. For further information, contact Jo Ann Jones, MAEF General Program Coordinator, School of Business, Georgia College, Milledgeville, GA 31061.

Call for papers. The **Association for the Advancement of Policy, Research and Development in the Third World** will hold its eighth annual conference on "International Science and Technology" November 20–24, 1988, in Castries, St. Lucia. Send abstracts, request for conference information, and any other communications to Dr. Shah M. Mehrabi, Vice President, AAPRD, Department of Economics, Mary Washington College, Fredericksburg, VA 22401-5358. Phone: 703-899-4092.

Call for papers. **The International Association for Energy Economics** will hold its annual North American Conference in Houston, TX, October 30 to November 2, 1988. The theme is "Competition, Profitability and Policy in the Energy Industries." The meeting, to focus on the future impact of the continuing restructuring of the energy industries, will include papers and discussions of the financial, competitive, strategic, regulatory, and policy aspects of that restructuring. To present papers, contact: Mr. Guy Caruso, Program Chairman, 1988 North American Meeting, 1133 15th St., N.W., Suite 1000, Washington, DC 20005. Phone: 202-586-2443. To attend, contact: Mrs. Joan Walsh Cassedy, Executive Director, International Association for Energy Economics, 1133 15th St., N.W., Suite 1000, Washington, DC 20005. Phone: 202-293-5913.

The **Stetson School of Business and Economics of Mercer University** is sponsoring a session at the Southern Humanities Conference in Macon, GA, February 16 to 18, 1989. Persons interested in presenting a paper or serving as a discussant should submit an abstract by October 31, 1988, to William S. Mounts, Jr., Stetson School of Business and Economics, Mercer University, Macon, GA 31207. Papers should link economics and the humanities in topics like: rhetoric of economics or the moral philosophy of modern economics.

Call for papers. ***Land Economics*** will publish a special issue, "Private Markets, Public Decisions: An Assessment of Local Land-Use Controls for the 1990s." Original papers with empirical, theoretical, institutional and law-and-economics perspectives are solicited. Regular referee procedures will be employed. Submit papers to William Fischel, Special Editor, *Land Economics*, 427 Lorch Street, Madison, WI 53706. Direct inquiries to William Fischel, Department of Economics, Dartmouth College, Hanover, NH 03755. Phone: 603-646-2940.

Call for papers. The **1989 Marxist Scholars Conference** will be held March 16–19, 1989, at the University of Louisville, KY. Completed papers or proposals for panels and workshops are due November 1, 1988. Send one copy of paper or proposal to each of the following: Professor Rinda Lundstrom, Department of Theater Arts, Theatre Annex, University of Louisville, Louisville, KY 40292; Professor Daniel Todes, Institute of History of Medicine, Johns Hopkins University, 1900 E. Monument Drive, Baltimore, MD 21205. The conference is sponsored by the Marxist Educational Press and papers will be considered for publication in the MEP journal *Nature, Society, and Thought*.

The third annual session of the **Mediterranean Summer School of Industrial Economics** will be held in Corsica at the Institution of Scientific Studies, September 26–October 1, 1988. The Mediterranean School promotes the spreading of knowledge and works in the field of industrial organization. For information write: P. M. Romani–A. Torre, Latapses-CNRS, 1 Avenue A. Einstein, Sophia-Antipolis, 06560 Valbonne, France.

The **National Humanities Center** offers 35–40 fellowships for advanced study in the liberal arts. Social and natural scientists or professionals whose work has a humanistic dimension can apply as well as scholars from any nation. Applicants must hold doctorates or have equivalent professional accomplishments. Fellows must work at the Center. Fellowships are for the academic year, September through May. A few may be available for a single semester. Application for the academic year 1989–1990 is October 15, 1988. Write to: Kent Mullikin, Assistant Director, National Humanities Center, 7 Alexander Drive, P.O. Box 12256, Research Triangle Park, NC 27709-2256.

Harvard Law School offers four or five Liberal Arts Fellowships to college and university teachers in the arts and sciences for the academic year 1989–90 at Harvard Law School. The purpose is to study fundamental techniques, concepts and aims of law so they will be better able to use legal materials and insights relevant to their disciplines. The year will not count toward a degree. The fellowship grant covers tuition and health fees only, plus office space. Applications should include a biographical resume (including academic record and list of publications), a statement noting what the applicant hopes to achieve through the year of study, and two letters of recommendation (mailed to the Chair directly from the referees). There is no special application form. Applications should be sent to: The Chair, Committee on Liberal Arts Fellowships, Harvard Law School, Cambridge, MA 02138. Applications for 1989–90 should be complete by January 15, 1989.

The **Council on International and Public Affairs** has announced an essay contest on Universal Capital Ownership with prizes of $5,000,

$2,500 and $1,000. Essays are limited to 5,000 words and should address the topic: How can broadened stock ownership be used to alleviate the problems of employment, under-employment, and poverty-level wages? The deadline is December 31, 1988. For information write (and include a stamped self-addressed envelope) to: USOP Essay Contest, Council on International and Public Affairs, 777 United Nations Plaza, New York, NY 10017.

Deaths

Henry Simon Bloch, managing director, E. M. Warburg Pincus & Company, adjunct professor emeritus of law and international relations, Columbia University, February 28, 1988.

H. Laurence Miller, emeritus professor of economics, University of Hawaii at Manoa, March 2, 1988.

George R. Morrison, professor of economics, Graduate School of Management, Rutgers University, December 15, 1987,

New Appointments

Gerald E. Auten, Bowling Green State University: visiting Financial Economist, Office of Tax Analysis, U.S. Treasury Department, 1987–1988.

Badi H. Baltagi: professor, Department of Economics, Texas A & M University, September 1, 1988.

Andrew F. Brimmer: Wilmer D. Barrett Professor of Economics, University of Massachusetts, September 1, 1987.

Frank J. Chaloupka IV: assistant professor, Department of Economics, University of Illinois at Chicago, September 1, 1988.

Rafael Del Villar: assistant professor, Economics Department, Texas A & M University, September 1, 1988.

Burkhard Drees: assistant professor, Economics Department, Texas A & M University, September 1, 1988.

Gerald Epstein, New School for Social Research: assistant professor, University of Massachusetts, September 1, 1987.

Matthew S. Goldberg, Center for Naval Analysis: economist, The Rand Corporation, September 1987.

Bruce R. Kingma: assistant professor, Economics Department, Texas A & M University, September 1, 1988.

John Mullahy, Yale University: assistant professor, Economics Department, Trinity College, September 1, 1988.

William S. Neilson: assistant professor, Economics Department,. Texas A & M University, September 1, 1988.

Elizabeth L. Rankin. assistant professor, Department of Economics, Centenary College, September 1, 1988.

John Rogers: assistant professor, Department of Economics, Southern Illinois University at Edwardsville, September 1987.

Jeannette A. Roskamp: Massachusetts Institute of Technology: associate economist, The Rand Corporation, March 1987.

Lisa L. Saunders: assistant professor, Department of Economics, University of Massachusets, September 1, 1987.

Harold Winter: assistant professor, Economics Department, Texas A & M University, September 1, 1988.

To Departmental Representatives and Executive Officers

When submitting information for inclusion in "Notes," please observe the following guidelines:

Calls for papers, notice of professional meetings, and other announcements of interest to economists should be submitted in one paragraph that contains all relevant information.

News of individual members should be labeled as to category: (1) deaths; (2) retirements; (3) foreign scholars visiting the US or Canada; (4) promotions; (5) administrative appointments; (6) new appointments; (7) leaves for special appointments; (8) resignations; (9) miscellaneous. Give individual's name, present place of employment, new title, new institution, and date on which change is to occur.

Deadlines for each issue are: Spring, January 1; Summer, April 1; Fall, July 1; Winter, October 1.

Please send all information to the *Journal of Economic Perspectives*, Department of Economics, Stanford University, Stanford, CA 94305-6072. We reserve the right to edit material received.

AMERICAN ECONOMIC ASSOCIATION

1988 ANNUAL MEMBERSHIP RATES

Membership includes:

—a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

● Regular members with annual incomes of $30,000 or less	$38 50
● Regular members with annual incomes above $30,000 but no more than $40,000	$46.20
● Regular members with annual incomes above $40,000	$53.90
● Junior members (available to registered students for three years only) Student status must be certified by your major professor or school registrar	$19.25
● In Countries other than the U.S.A., Add $16.00 to cover postage.	
● Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association)	$7.70

Please enter my subscription for the following period:

☐ **Jan.-Dec.** ☐ **April-March** ☐ **July-June** ☐ **Oct.-Sept.**

First Name and Initial ____________ **Last Name** ____________ **Suffix** ____________

Address Line 1 ____________

Address Line 2 ____________

City ____________

State or Country ____________ **Zip/Postal Code** ____________

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

☐ ☐

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____________

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____________

____

Authorized Signature ____________

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 3-7

Symposium on the Slowdown in Productivity Growth

Stanley Fischer

The growth slowdown that began in the late 1960s or early 1970s is the most significant macroeconomic development of the last two decades. Table 1 presents the basic facts for the big seven members of the OECD. The significance of the slowdown derives from the power of compound interest: when per capita income is growing at 4 percent it doubles every 18 years; when it grows at only 2 percent it doubles every 35 years. The permanent income, defined as the annuity value of expected income, of a 25 year old German citizen basing her expectations of income growth on average performance in the period 1955-73 would be 80 percent above her current income; that of a 25 year old basing her expectations on the growth rate of income in the 1973-1986 period would be only 22 percent above her current income.[1] Equivalently, taking a generation as 33 years, children in Italy who might have expected to start out with a standard of living nearly five times the level at which their parents began, would on the basis of the lower output growth from 1973 to 1986 expect to start out at less than double the material standard at which their parents began.

The four papers in this symposium address the question of the causes of the growth slowdown and, to a lesser extent, what might be done about it. The classic empirical result in this area (for which, among other results, Robert Solow received the 1987 Nobel Prize in economic science) is that technological progress accounts for about half the growth of per capita output. Correspondingly, most of the growth

---

[1] The assumptions are that the real discount rate remains constant at 4 percent, that the person concerned is 25 years old and expects income to grow at the same rate as aggregate GNP per capita, expects to retire at 65, and die at 80.

■ *Stanley Fischer is Vice President for Development Economics, and Chief Economist, The World Bank, Washington, D.C.*

Table 1
Percentage growth of per capita GNP

	1955–73	*1973–86*
United States	2.0	1.3
Japan	8.8	3.1
Germany	4.2	2.0
United Kingdom	2.5	0.9
Italy	4.9	2.0
France	4.6	1.9
Canada	3.0	2.0

Source: International Financial Statistics. Data are for GDP in some cases.

slowdown (in per capita income) can be traced to a slowdown of productivity growth —though, as we shall see, there is some disagreement over both the role of productivity growth in economic growth and its contribution to the growth slowdown. Table 2 give estimates of U.S. productivity growth, defined as output per hour in the nonfarm business sector.

The data in Table 2 are what is most frequently used to describe the slowdown in productivity growth. Perhaps more useful are data on multifactor productivity growth, the residual after subtracting the growth attributable to increases in inputs, derived by weighting the growth rate of each factor by its share of output. These data are now calculated by the Bureau of Labor Statistics and show a pattern similar to that of Table 2. Because the contribution of capital, which has generally grown faster than labor, is removed, multifactor productivity growth is typically below that of output per hour. Multifactor productivity growth in the nonfarm business sector over the period 1957 to 1973 averaged 1.5 percent; in the period since 1973 it has been 0.5 percent.

There are several well-known suspects in the search for the causes of the productivity growth slowdown.[2] It it generally recognized that because the change was worldwide, the explanation for the post-1973 slowdown is unlikely to lie in the special circumstances of a single country.[3] Timing alone gives the lead role to the energy price increase of 1973. Another possibility is a slowdown in the rate of production of relevant knowledge: in this interpretation the extraordinary growth of 1945–1973 represents a catch-up from the Great Depression and technical advances of World War II, which has been replaced by a return to the longer-run average rate of productivity growth. Other candidates include a decline in the rate of increase of the skill of the labor force, a decline in the rate of investment, the shift towards the service economy, and the increasing role of government in the economy.

What do the authors in this symposium tell us? To the extent that any one explanation is favored, it is that the oil price shock is to blame. Dale Jorgenson

---

[2] For a basic growth accounting analysis of the slowdown, see Denison (1985), Chapter 3. Two very useful recent contributions by Baily (1986) and Berndt and Wood (1987) identify and evaluate the roles of many of the suspects. They should be considered a supplement to the papers in this symposium.

[3] Baily (1986) shows that the slowdown after 1968 was particular to the United States.

Table 2
U.S. annual productivity growth, 1947–1986 (percent)

1947–1955	*1955–1968*	*1968–1973*	*1973–1980*	*1980–1986*
2.9%	2.5%	1.5%	0.4%	1.1%

Source: Economic Report of the President, 1987, Table B-43.

presents a detailed calculation of the sectoral impacts of the oil price shock on productivity growth, while Mancur Olson explains why economic systems may have become increasingly incapable of responding flexibly to a shock such as the increase in the price of oil. Zvi Griliches examines direct evidence on changes in the rate of creation of new knowledge. Michael Boskin's contribution is directed less at the causes of the productivity growth slowdown and more at the question of how tax policies affect productivity and productivity growth. We briefly discuss the contributions in the order in which they appear.

In a characteristically balanced assessment of the role of research and development in the growth slowdown, Zvi Griliches is reluctant to point a finger of blame, though there is evidence in favor of a slowdown in the rate of creation of knowledge. In fact, patenting activity slowed down in the 1960s. However, Griliches is skeptical of the importance of this factor. He emphasizes the possibility that measurement problems account for part of the slowdown: any economist running cross-sectional regressions on a portable computer seven miles above the Atlantic, removing her Walkman only long enough to call her broker to execute a trade in Hong Kong, finds it difficult to think that life is not much better than it was, but it is doubtful whether much of this shows up in the GNP accounts.[4] The measurement problem may well be very serious, and further work on this is to be expected. Beyond measurement problems, Griliches is inclined to attribute the slowdown in productivity growth to reduced levels of aggregate demand, which operate in part by reducing capacity utilization.

On the basis of disaggregated production functions, Jorgenson argues that the productivity growth slowdown 1973–1979 can be attributed largely to the rise in the price of energy.[5] The analysis is conducted at a sectoral level. In principle, an increase

[4] Martin Baily (1986) presents some evidence that the office revolution has mostly generated more paperwork. He does not enter into the question of the social value of the increased output of economics papers.

[5] Although all the authors in the Symposium agree on the broad fact of a decline in productivity growth, Jorgenson's productivity growth rates are lower than those based on BLS estimates. Part of the difference arises from Jorgenson's use of disaggregated data, part from differences in measurement of inputs, and part because Jorgenson estimates production functions for gross output, with intermediate inputs treated explicitly, while other estimates are for value added production functions. Berndt and Fuss (1986) present alternative estimates of productivity growth; over the period 1948–81 they estimate multifactor productivity growth in U.S. manufacturing to be 1.54 percent by the value added measure and 0.54 percent in the gross output production function.

in the price of energy could increase, decrease or leave unchanged the rate of change of productivity in any given industry. Jorgenson estimates that for most industries, increases in the energy input augment technical progress. Thus, for most industries the decline in the use of energy reduced the growth rate of productivity. Presumably rates of increase of productivity will increase as a result of recent lower energy prices.

Mancur Olson's contribution takes a broader political economy view of the growth process. He starts by restating his well-known thesis that interest groups gradually encrust an economy's performance, and that the more stable the country, the slower its growth. (Perhaps this applies too to the Soviet Union.) Being big think and very general, the thesis is subject to detailed criticisms. Olson's purpose in this paper is to apply his thesis to the growth slowdown. Since the slowdown was worldwide, he cannot rely on the simple form of the hypothesis in which the more stable countries continue to grow more slowly. Rather, he argues now that the interest groups not only interfere with the growth process, but also slow the reaction of the economy to real shocks. Combined with an argument based on hysteresis, this enables Olson to account in a general way for the aggregate effects of the oil shock. However, the cutting power of the hypothesis is unclear. For instance, why was Japan hit so hard by the first oil shock but not the second? Why was Germany hit less hard by the first oil shock than the second?

Michael Boskin addresses the role of the tax system in affecting economic growth. He has two main conclusions. First, although it was for a while fashionable to emphasize the potential distortions arising from the misallocation of capital due to different rates of taxation, those costs are in fact quite low. Rather, Boskin argues, the main distortion caused by the present tax system is that it discriminates against saving. In accordance with estimates that the supply of saving is quite sensitive to its rate of return, he argues that major benefits for growth would flow from a move to a consumption tax. Second, although the ERTA-TEFRA tax acts of the early 1980s contributed to growth, their positive contribution was overshadowed by the adverse effects of the aggregate budget deficit on growth. Boskin emphasizes increased productivity growth over the 1981–1987 period compared with the previous eight years. However, this comparison is quite sensitive to the division of the period.

Where do we stand now? We have in this symposium a sample of views, with some emphasis on the increased price of energy. Although Jorgenson provides clear estimates of the impact of energy prices, the overall impression is that economists are as yet unable to pin down the relative contributions of the potential causes of the productivity slowdown. However, it is noteworthy that the explanations that focus on energy prices and the Griliches hypothesis that a lack of aggregate demand accounts for the slowdown both imply a revival of productivity growth, at least in the United States, where the unemployment rate is for the first time in a decade close to its natural level.

■ *I am grateful to Ernst Berndt and Zvi Griliches for discussions, and to Michael Boskin and Dale Jorgenson for comments.*

References

Baily, Martin N., "What Has Happened to Productivity Growth?" *Science*, Oct. 24, 1986, 443–451.

Berndt, Ernst R., and Melvyn A. Fuss, "Productivity Measurement with Adjustments for Variations in Capacity Utilization and Other Forms of Temporary Equilibrium," *Journal of Econometrics*, 1986, *33*, 7–29.

Berndt, Ernst R., and David O. Wood, "Energy Price Shocks and Productivity Growth: A Survey." In Gordon, R. L., *et. al*, eds., *Energy: Markets and Regulations, Essays in Honor of M.A. Adelman*. Cambridge, MA: MIT Press, 1987.

Denison, Edward F., *Trends in American Economic Growth, 1929–1982*. Washington, D.C.: Brookings Institution, 1985.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—pages 9–21

Productivity Puzzles and R&D: Another Nonexplanation

Zvi Griliches

A number of facts are not in dispute. Around 1974–75, output growth and associated productivity measures dropped sharply in the United States and in most other western industrialized nations and continued at rather low rates for most of the rest of the 1970s. This drop in productivity growth was actually larger, in absolute terms, in some other countries (such as Canada, Japan, and Sweden) than in the United States, making explanations for these events which rely heavily on specifically U.S.-based causes or arguments somewhat less plausible. Whether the productivity slowdown has come to an end in the mid-1980s is still very much in dispute. So too is the possibility that the slowdown actually started earlier, in the mid-1960s, as is implied by some of the more inclusive, multifactor productivity measures.

During the late 1960s the growth in research and development (R&D) investment slowed down markedly (in constant dollars) and did not really recover until the late 1970s. Basic research was especially hard hit, showing a substantial absolute decline during the same period. Whether this slowdown in the investment in new technologies can account for the observed productivity slowdown is a fascinating question. I shall argue below that it cannot, at least not yet, since its effects take a long time to work themselves through the innovation and diffusion processes. The oil price hikes of the early and late 1970s and their macro-consequences are, therefore, the most likely direct causes of these pervasive declines in the growth rate of productivity. I will begin, however, by reviewing first some relevant aspects of the productivity slowdown and some of the explanations for it proffered in the literature.

■ *Professor Zvi Griliches is Paul M. Warburg Professor Economics, Harvard University and Director of the Productivity Program at the National Bureau of Economic Research, both in Cambridge, Massachusetts.*

The Slowdown and Its Pervasiveness

There are two major measures of productivity growth used routinely by economists: output-per-man-hour and "total factor" or "multifactor" or "residual" productivity. The first measures productivity growth by the difference in rate of growth of an output index based on value-added in constant prices and an index of man-hours worked in a country or industry. The second measure subtracts from the first an estimate of the contribution of physical capital to productivity growth, based on the growth in the capital-labor ratio weighted by the smoothed share of capital inputs in factor payments. Both measures suffer from difficulties in computing correctly an index based on value-added "real" output in a world of changing commodities and services and in measuring the quantity and quality of labor service. The second measure suffers from the additional problem of how to measure capital in general and its utilization rate, in particular. I shall return to the discussion of some of these difficulties below. In the meantime, since these are the numbers that we have and that most people are looking at, I will proceed as if they actually mean what they are supposed to mean.

Table 1 uses the first measure, output per man-hour, to document the two major facts about the recent slowdown in productivity growth: its worldwide nature and 1974–75 as the date of its onset. Table 2 uses the second measure, total factor

Table 1

Output per Man-Hour in Manufacturing: Average Annual Growth Rates Selected Countries and Periods

Country	*1950–60* *(1)*	*1960–73* *(2)*	*1973–79* *(3)*	*1979–86* *(4)*	*Deceleration 1973–79 minus 1960–73* *(5) = (3)–(2)*
U.S.	2.0	3.2	1.4	3.1	−1.8
Canada	3.8	4.5	2.1	1.4	−2.4
Japan	9.5	10.3	5.5	5.6	−4.8
Belgium	n.a.	6.9	6.2	5.3[a]	−0.7
Denmark	2.8	6.4	4.2	1.3	−2.2
France	2.8	6.5	5.0	3.6	−1.5
Germany	7.4	5.8	4.3	2.8	−1.5
Italy	5.7	7.3	3.3	3.3	−4.0
Netherlands	4.7	7.4	5.5	4.4[a]	−1.9
Norway	3.4	4.3	2.1	1.9	−2.2
Sweden	3.4	6.4	2.6	3.0	−3.8
U.K.	2.1	4.3	1.1	4.4	−3.2

[a] —1979–85 n.a.—not available

Source: Bureau of Labor Statistics, "International Comparisons of Manufacturing Productivity and Labor Costs Trends, 1986," USDL 87–237.

Table 2

Estimates of Total Factor Productivity: Average Annual Rates of Growth (in percent) in the U.S., by sectors and periods.

Sector	*Periods* *1949–1966*	 *1967–1973*	 *1974–1985*
Total Private Industry	1.3	0.4	0.1
Agriculture, Forestry, & Fisheries	−1.8	0.1	1.4
Mining	0.5	1.9	−4.3
Construction	1.7	-5.1	−1.1
Manufacturing	2.0	1.8	1.4
Transportation	1.0	1.6	0.1
Communication	2.7	3.0	2.3
Utilities	3.7	2.4	1.2
Wholesale Trade	0.8	0.8	−0.5
Retail Trade	0.6	−0.0	−0.4
Finance, Insurance, etc.	−0.3	−0.9	−0.6
Services	−0.6	−0.2	0.1

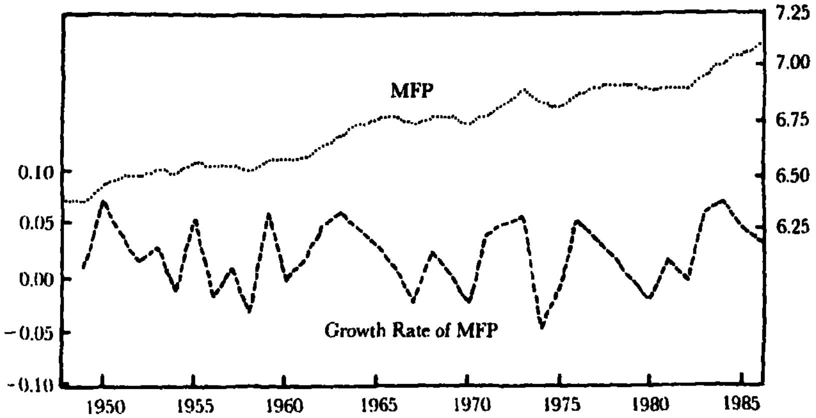
Source: From Shapiro (1987). Based on BEA estimates of GNP, compensation, hours of work and capital stock by industry. Approximate TFP computed as the difference in the rate of growth of the output-capital ratio and the labor share weighted growth in the total hours to capital ratio. Share of labor compensation in GNP averaged over t and $t-1$. TFP growth $= (y - c) + \bar{s}_l(l - c)$, where $\bar{s}_l(t) = [s_l(t) + s_l(t-1)]/2$ and s_l is the share of labor payments in total factor payments.

productivity, to present similar data for different sectors of the U.S. economy. The story is somewhat more confused here. In several sectors, especially construction, the productivity slowdown started significantly earlier, in the mid-1960s. In others, such as communication, productivity has hardly slowed down at all. Some sectors, especially manufacturing (see Figure 1), appear to be recovering their pre-1974 growth rates, raising the hope that the overall slowdown may also come to an end in the not too distant future.

A look at Table 2 raises a variety of puzzles and some doubt about the ability of economists to measure output and input adequately over time and across industries, sectors, and countries. Is it reasonable to conclude that total factor productivity in mining and construction is significantly lower today than had been its level in the early 1950s and even earlier, and that there has been no growth in productivity during the whole post-World War II period in the financial and other service sectors? This is not really credible and thus makes one doubt some of the other numbers that are not obviously strange at first glance. There are indeed severe measurement problems associated with concept of productivity in its various incarnations, a topic to which I shall keep returning.

The question of whether the observed slowdown is a "cyclical" fluctuation or represents a serious break in the underlying historical trend can also be answered differently depending on the industry and level of aggregation chosen. Figure 1 plots the level and rate of growth of total factor productivity in U.S. manufacturing from

Fig. 1. Total factor productivity: U.S. manufacturing 1949–1986, levels and growth rates, 1949–1986 (*Log scale*)



Source: U.S. Department of Labor, "Multifactor Productivity Measures 1986"

1949 through 1986. The figure shows a large amount of year to year fluctuation, but no visible evidence of a consistent decline in its annual growth rate. A regression of the growth numbers on a trend variable yields a small positive but statistically insignificant coefficient, confirming the visual impression gathered from Figure 1 of no significant decline in the rate of growth of total factor productivity in manufacturing over the whole of this period. Somewhat different data, taken from Shapiro (1987), lead to the same conclusion for the 1951–85 period. If one chooses a specific date a priori for the break, say 1974, one does indeed get a striking difference between the pre-1974 (1951–73) average rate of growth in total factor productivity (1.8 percent per year with a standard deviation of 3.0 percent) and the post-1974 period (1974–82): 0.2 percent with a standard deviation of 3.0 percent), but this difference is still not statistically significant at conventional significance levels.

The story is different for more economy-wide measures of productivity growth, such as the total private business sector, where there is a distinct and statistically significant decline in the average rate of growth since the mid-1970s. This is the result of what is already visible in Table 2, the much sharper and earlier productivity slowdowns outside of the manufacturing sector. Since I believe that output and hence productivity are better measured in manufacturing than in most of the other sectors of the economy (excluding agriculture, electric utilities and communications) and since the measurable effects of research and development are most likely to show up first in manufacturing (and agriculture and communications), I interpret these numbers to imply that there is little evidence of a systematic and permanent decline in the contribution of science and R&D to productivity growth during the last decade or two. Without a consistent long-term productivity slowdown in the sectors where the effects of science and R&D should be visible first and where the measurement issues are not absolutely hopeless, there is little point in developing and exploring various "end of the scientific frontier" stories. Nevertheless, I shall pursue this line of argument a bit further in the next section of this paper.

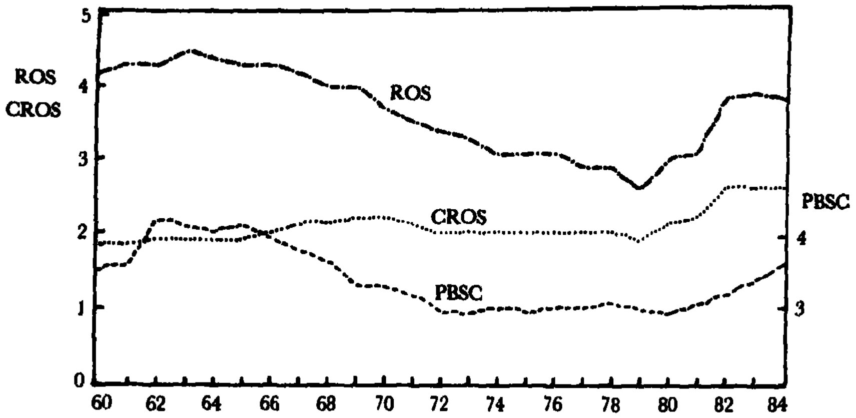
Connecting R&D to Productivity

Of the various explanations for the productivity slowdown only two have some verisimilitude and are worth further exploration: the rise in energy and material prices and the possible exhaustion of the wellsprings of technological progress. These and many other explanations have been reviewed repeatedly recently—see especially the contributions of Baily (1986), Denison (1986), Jorgenson (1986), Jorgenson, Gollop and Fraumeini (1988), Kendrick (1983) and Maddison (1987)—without any clear consensus emerging about the perpetrator of the crime or even the actual date and locus of the event itself. Energy price rises did occur at the right time and also worldwide and hence are the prime suspect in this case, but the exact mechanism of their impact is still in dispute. Energy use in industry did not fall significantly, nor did the price rises induce major resource-using technical changes quickly. Nevertheless, productivity growth slowed down almost immediately and very widely, in a variety of countries and industries. The mechanism for this aspect of the slowdown must lie, therefore, elsewhere: in the workings of the international macroeconomy rather than in the direct impact of energy prices on production technology at the micro enterprise or industry level.

Figure 2, which plots several different measures of U.S. research and development spending from 1960 on, raises the possibility that the observed decline in R&D in the mid-1960s could have contributed to the persistence of the productivity slowdown in the 1970s. Given the lags that are associated with the impact of R&D on productivity, the timing is about right. However, this view has several potential difficulties. First, the R&D slowdown was much less pronounced in other Western countries and in Japan, in spite of the even larger declines in productivity growth in those other countries. Second, most of the slowdown in the growth of R&D occurred in its governmentally supported components (which equals the difference between the ROS and CROS lines in Figure 2), while most of the available evidence argues that this type of R&D has had only a relatively small role in generating productivity growth in industry (Lichtenberg, 1984; Terleckyj, 1974; and others). Third, a direct computation of the potential contribution of the slowdown in R&D growth to the explanation of productivity growth results in a rather small number. Before this last point is taken at face value, however, I need to explain briefly how such computations are actually made.

A common and popular approach to the incorporation of R&D into growth accounting schemes is to treat it as another type of capital which is to be added to the list of aggregate production function variables (Griliches, 1979; Mansfield, 1980; Scherer, 1982; Terleckyj, 1974). Passing over a long list of conceptual and empirical

Fig. 2. R&D expenditures in U.S. industry, 1960–1984



ROS—Ratio of R&D to sales, in percent
CROS—Ratio of company financed R&D to total sales
PBSC—Basic research as a percent of total R&D expenditures
Source: NSF (1986)

problems associated with the concept of R&D "capital," one can write an expanded version of the standard output growth equation as

$$y = \alpha c + \beta l + \gamma k + t,$$

where y, c, l, k, and t are the rates of growth of output, physical capital, labor input, R&D "capital" (k = "knowledge"), and disembodied technical change respectively, and γ is the elasticity of output with respect to R&D capital. Since $\gamma = (\Delta Y/\Delta K)(K/Y)$, while $k = (\Delta K/\Delta t)/K$, one can "simplify" the γk term and rewrite this equation as

$$y = \alpha c + \beta l + \rho(R/Y) + t$$

where $\rho = \Delta Y/\Delta K$ is the marginal product of R&D capital or the gross rate of return to it and $R = \Delta K/\Delta t$ is the net investment in this "capital." This second equation expresses output growth as a function of R&D "intensity" (R/Y) and is relatively easy to discuss and implement provided one is willing to assume that there is no or very little depreciation in this type of capital. Then, and this is what is done most often, the observed ratio of R&D to GNP or sales is identified with the unobservable "net" investment ratio allowing the analysis to proceed on the basis of the available data. Given an observable R/Y, all (!) one needs for a back of the envelope calculation is an estimate of ρ, the rate of return to R&D. Since one does not expect that social returns to R&D are the same as private returns, one cannot use observed income and balance sheet data to compute the economy- or industry-wide ρ. Nor can one use them to compute the private ρ directly, since it is not clear how one allocates total private returns to different types of "capital." Rather, one has to base such estimates on econometric or case study calculations (Griliches, 1958, 1973; Mansfield et al, 1977).

In this context, the assumption that knowledge does not depreciate is not too bad if one is interested in time series comparisons within a well-defined industry or aggregate and the total stock of R&D capital is changing only slowly. Then the difference between changes in gross and net investment is unlikely to be very large. This, of course, need not be true for cross-industry comparisons or for periods where one might suspect a significant obsolescence of the previously accumulated but now irrelevant R&D capital; for example, the situation following the rise of OPEC, with past R&D results having been based on the assumption of continued low oil prices and hence subject to significant obsolescence.

Versions of the first equation have been estimated by Griliches (1980a, 1986) and by Griliches and Mairesse (1984) using time series data on individual U.S. companies, and by Griliches (1973, 1980b) and others using industry level data. The estimated elasticity of output with respect to R&D capital tends to lie between .06 and 0.1. These elasticity estimates are not sensitive to whether one uses total R&D or only company-financed R&D based capital stock measures, though there are other indications in the data that government financed R&D produces less benefit than privately financed R&D. Versions of the second equation have been estimated by Griliches (1986), Clark and Griliches (1984), Link (1981a, 1981b), and Mansfield (1965), among others, using firm or line-of-business level data, and by Griliches (1976), Griliches and Lichtenberg (1984a, 1984b), Scherer (1982, 1984), Terleckyj (1974) and others using industry aggregates. The estimated rate of return to R&D lies mainly between 0.2 and 0.5, with most of the recent estimates falling in the lower part of this range. While the presence of spillovers would make one expect the industry level coefficients to be higher than those estimated at the firm level, the econometric estimates do not show this in any convincing fashion. For evidence of spillovers one must look elsewhere (Jaffe, 1986; Scherer, 1982) where alternative measures of R&D expenditures by "other" firms or industries, weighted by technological distance or research "intention" weights are shown to contribute significantly to productivity growth in the receiving industries. But even here, the estimated effects are not very large. There is, however, strong evidence that the yield on federally financed R&D investment (its contribution to firm and industry-wide productivity growth) is significantly lower than on company financed projects (see also Lichtenberg, 1986).

It turns out that coefficients of this size, even though they imply a substantive and statistically significant role for R&D in the productivity growth story, are not large enough to account for more than a modest part of the observed slowdown in the growth of productivity. This conclusion traces back to the fact that the more relevant privately financed R&D investments did not really decline during this period, as Figure 2 shows. Even looking at the growth rate of total R&D stocks which declined by about 3 percent between the mid-1960s and the mid-1970s (Griliches, 1980b) and using the upper range of the estimated elasticities of output with respect to R&D capital, 0.1, one can account for only about 15 percent of the observed slowdown in *total factor productivity growth in manufacturing since 1974. One gets rather similar results based on estimates of the second equation. Using a $\rho = 0.3$ and the average decline in the total R&D to sales ratio in manufacturing of minus 1.1 percent between the 1960s and the 1970s, one reaches essentially the same conclusion: the slowdown in R&D investment can account for only a small fraction of the productivity slowdown. Using company financed R&D measures or making a downward adjustment to the contribution of federally financed R&D would only reduce this estimate further.*

Have R&D Opportunities Been Exhausted?

There are a number of ways of trying to make something more out of the R&D story as it pertains to the observed slowdown in productivity growth. One could try asking whether the efficiency with which R&D is translated into productivity improvements has declined over time. Has the economy run into diminishing returns due to the exhaustion of new invention opportunities? My tentative answer is no. There is no convincing evidence of such an exhaustion in the data or in the technological news in the daily newspapers. One way to test this question is to look for changes over time in the estimated R&D coefficients, and here the available data do not yield a consistent story. In an earlier study based on the first equation (Griliches, 1980b), I thought that I had detected a decline in γ between the early sixties and early seventies. But using microdata, I found subsequently higher coefficients for 1966–77 than had been estimated earlier for the 1957–65 period using similar methods (Griliches, 1980a, 1986). Using the approach of the second equation, higher estimates of ρ for the 1970s were also found by Griliches and Lichtenberg (1984a, 1984b) and Scherer (1982).

But what about the evidence of a decline in "inventiveness" implied by the numbers on innovations collected by Baily and Chakrabarti (1985) and the evidence on the decline in the number of patents granted to U.S. inventors in the 1970s? The main question here is one of cause and effect. It is well known that such measures of inventive activity are heavily influenced by economic conditions and prospects (Schmookler, 1966). Patent applications, for example, fell sharply in the 1930s, recovered in the late 1940s, and were essentially flat from the early 1950s through the early 1960s. They rose sharply in the late 1960s and then basically flattened out throughout the 1970s, starting to rise again a bit in the 1980s. Most of the rise in patent applications in the late 1960s and early 1970s came from a major rise in applications from abroad which effectively "crowded out" a significant number of U.S. patents, given a relatively fixed resource level available to the U.S. Patent Office for processing such applications.

The question of whether R&D opportunities are becoming exhausted is clearly important and requires more study and better data. But based on the available data, I remain unconvinced that a turning point has occurred in the recent past and that the world now faces significantly less promising research opportunities than it did several decades ago. I interpret most of the proffered evidence as reflecting the impact of reduced aggregate demand and less favorable economic prospects for inventive activity in the late 1970s, rather than as the result of the technological springs running dry. This does not mean that one can expect high rates of return from R&D forever or continued high rates of disembodied technical change into the indefinite future, only that the world did not end, in this regard, in 1974.

The other possibility is that the returns to R&D did not decline in the 1970s but that the large changes in relative prices, especially of energy, caused an obsolescence of much previously valuable economic knowledge and an associated "write-down" of the stocks of R&D capital. This is similar to Baily's (1981) point about physical capital obsolescence and to Jorgenson's (1986) interpretation of the impact of rising energy prices on the aggregate rate of technical change. I believe that something did happen along these lines, but that the magnitude was not large enough to account for much of the observed productivity slowdown. First, a decline in the previously available stock of R&D capital should have provided a significant stimulus for an increase in the rate of R&D expenditures, but no such rise is observed until the late 1970s and early 1980s. That rise is most likely associated with the rising aggregate defense expenditures (Lichtenberg, 1984) and possibly also with some losses in R&D capital induced by the successes of our international competitors. But that is a different and later story.

Nor was there enough change in energy use in the mid-1970s or in the technologies used to produce most industrial products to allow the interpretation that the rise in oil prices led to a move to a different, less favorable (at the old relative prices) location along the productions possibilities frontier. Certainly, the economy moved away from the frontier and did not utilize its resources fully during this period, but a recession should not be interpreted as implying that the frontier itself shifted downward. When society is not at the frontier, we cannot really tell what is happening to it.

The Difficulties of Assessing the R&D Contributions

Before I leave the research and development story entirely I want to sound a warning note about potential misinterpretations. The fact that one cannot account for much of the productivity slowdown by looking at what happened to R&D does not mean that R&D is not important to the productivity story, only that R&D did not change enough during this particular period to cause the observed change in productivity. Moreover, it is likely that the type of study described above seriously underestimates the contribution of R&D to economic growth. This is true for at least two reasons: an inability of researchers to measure productivity correctly, especially in R&D intensive sectors and in services (which have been growing relative to the rest of the economy), and our inability to estimate correctly the magnitude of R&D spillovers from one firm, industry, or country to another.

The national income accounts, as currently constructed, do not reflect major components of the "product" of R&D and science and hence cannot serve as adequate measures of it. Most of the industrial R&D in the United States has been and is being spent on defense and space exploration connected projects. Its product is "sold" to the public sector and is measured, by accounting convention, by its costs. This implies a zero contribution to measured productivity growth except, perhaps, for its spillover effects on other products and industries. Whether moon landings are successful or not, their contribution to GNP in constant prices is about the same, since they are measured by their inputs and not by some independent measure of the success of the venture. Similarly, public expenditures on health research which lower the incidence of some disease will have little impact on national income. A reduction in morbidity would, to a first approximation, raise both GNP and hours worked, leaving output per man-hour largely unchanged.

Matters are only slightly better in considering industrial R&D that results in new products or improvements in the qualities of older products. The fraction of the direct social gain from such improvements that will show up in the national accounts as they are currently constructed depends on the ability of producers to appropriate the benefits of such innovations and on the treatment of new commodities in the official price indexes. If the producer does not have complete monopoly power over the invention, the price received will not reflect all of the social benefit, since part of it will be passed on to consumers in the form of lower prices per equivalent quality or performance unit. Whether the benefits of such price declines and quality improvements show up in the "real" GNP accounts depends on how they are treated in the construction of price deflators. If such improved products are used in turn as inputs in the production of other private products, these improvements will show up as a contribution to productivity growth in the industries that purchase them. Not being able to measure the improvements in farm equipment correctly, for example, the contribution of R&D in that industry will show up as an improvement in agricultural productivity and could be measured and captured there. But many new products are sold directly to consumers, or to industries where output is also difficult to measure, such as financial services, and hence the contribution of R&D invested in such products is unlikely to be captured fully in the conventional GNP accounts. For examples, consider the contribution to productivity of personal computers, computers sold to banks and brokerage houses, and jet engines. The fact that one can travel from the east to the west coast in less than half the time that it took two decades ago leaves hardly a trace in the national productivity accounts.

It is likely, therefore, that economists are seriously underestimating the true contribution of R&D to economic welfare by trying to measure it by its effects on standard productivity indexes. We are also unable as yet to capture well the spillover aspect of industrial R&D, the benefits that one firm or industry gets from the research of another. Much of the new knowledge created by R&D tends to seep out beyond the borders of where the research was done, leaving few traces as to where it leaked and whence it came. Recently there have been several interesting attempts to capture the impact of such spillovers and to model their spread but they have not yet yielded a convincing estimate of their global impact (Bresnahan, 1986; Bernstein and Nadiri, 1986; Jaffe, 1986; Scherer, 1982; Terleckyj, 1974). Our current understanding of this whole process is still seriously flawed. Moreover, without a major revision and extension of the national income accounts and the development of new data and methods for tracing the flow of ideas from one sector to another, researchers are unlikely to do much better in the near future.

One can show that R&D has been a significant contributor to past productivity growth and a good investment on average, that basic research has an even stronger effect on productivity growth and that the direct effects of federally financed R&D expenditures in industry are not as large as those of privately financed ones, but the size of the estimated effects may be seriously off, perhaps by an order of magnitude. While R&D is probably not the major culprit in the recent productivity slowdown and the associated erosion of our international competitiveness, it is hard to overestimate its importance for the long run growth of the world economy.

Returning to the Prime Suspect

What is then the culprit? Why has productivity grown so slowly in the last decade or so? My prime suspect remains the rise in energy prices and its macro consequences. It is not just that many industries had to face new prices, change the way they used their factors of production, and scrap much of their now unprofitable capacity, but also a long worldwide recession induced by the fall in real wealth caused by OPEC, by the fall in aggregate demand caused by the governments trying to control the resulting inflation, and the subsequent fall in U.S. exports and the increase in import competition in the early 1980s as the result of rising dollar exchange rates. These factors combined together to produce one of the longest worldwide recessions and growth slowdowns from which the world may not yet have emerged. The resulting prolonged periods of capacity underutilization in many industries is the proximate cause of much of the observed declines and slowdowns in productivity growth. This is also why one can find cross-country correlations between investment, price increases in energy and materials, and changes in productivity growth and find very little of the same at the detailed industry level within countries (Bruno and Sachs, 1985; Griliches and Mairesse, 1983). The main mechanism is the overall macro one and that is still where economists must go looking for a solution.

Of course, there may not be a single cause—one murderer. Perhaps it is more like the Murder on the Orient Express—they all did it!

From the longer run point of view there are still lingering doubts about the crime itself: perhaps the 1970s were not so abnormal after all. Maybe it is the inexplicably high growth rates in the 1950s and early 1960s that are the real puzzle. This thought, however, is a sad one. I still hope that when the world finds its way out of this worldwide growth recession, we will find that productivity will bounce back, that it has not left us motherless forever.

References

Baily, Martin N., "Capital Productivity and the Services of Capital and Labor," *Brookings Papers*, 1981, *1*, 1–50.

Baily, Martin N., "What Has Happened To Productivity Growth?" *Science*, 1986, *34*, 443–451.

Baily, Martin N., and Alok K. Chakrabarti, "Innovation and Productivity in U.S. Industry," *Brookings Papers*, 1985, *2*, 609–639.

Bernstein, Jeffrey I., and M. Ishaq Nadiri, "Research and Development and Intraindustry Spillovers: An Empirical Application of Dynamic Duality," NBER Working Paper No. 2002, July 1986.

Bresnahan, Timothy, "Measuring Spillovers from 'Technical Advance,'" *American Economic Review*, September 1986, *76*, 741–755.

Bruno, Michael, and Jeffrey Sachs, *Economics of Worldwide Stagflation*. Cambridge, MA: Harvard University Press, 1985.

Clark, Kim B., and Zvi Griliches, "Productivity Growth and R&D at the Business Level: Results from the PIMS Data Base." In Griliches, Zvi, ed., *R&D, Patents and Productivity*. Chicago: University of Chicago Press, 1984, 393–416.

Denison, E. F., *Trends in American Economic Growth, 1929–82*. Washington, D.C.: The Brookings Institution, 1986.

Griliches, Zvi, "Research Cost and Social Returns: Hybrid Corn and Related Innovations," *Journal of Political Economy*, October 1958, *66(5)*, 419–432.

Griliches, Zvi, "Research Expenditures and Growth Accounting." In Williams, B. R., ed., *Science and Technology in Economic Growth*. London: Macmillan, 1973, 59–95.

Griliches, Zvi, "Issues in Assessing the Contribution of R&D to Productivity Growth," *The Bell Journal of Economics*, Spring 1979, *10(1)*, 92–116.

Griliches, Zvi, "Returns to Research and Development Expenditures in the Private Sector." In Kendrick, J. W., and B. Vaccara, eds., *New Developments in Productivity Measurements*, NBER, Studies in Income and Wealth, Vol. 44, Chicago: University of Chicago Press, 1980a, 419–454.

Griliches, Zvi, "R&D and The Productivity Slowdown," *American Economic Review*, May 1980b, *70(2)*, 343–348.

Griliches, Zvi, "Productivity, R&D, and Basic Research at the Firm Level in the 1970s," *American Economic Review*, 1986, *76(1)*, 141–154.

Griliches, Zvi, and Frank Lichtenberg, "R&D and Productivity Growth at the Industry Level: Is There Still A Relationship?" In Griliches, Zvi, ed., *R&D, Patents, and Productivity*. Chicago: University of Chicago Press, 1984, 465–496.

Griliches, Zvi, and Frank Lichtenberg, "Interindustry Technology Flows and Productivity Growth: A Re-Examination," *Review of Economics and Statistics*, May 1984b, *LXVI(2)*, 325–329.

Griliches, Zvi, and Jacques Mairesse, "Comparing Productivity Growth: An Exploration of French and U.S. Industrial and Firm Data," *European Economic Review*, 1983, *21*, 89–119.

Griliches, Zvi, and Jacques Mairesse, "Productivity and R&D at the Firm Level." In Griliches, Zvi, ed., *R&D, Patents, and Productivity*. Chicago: University of Chicago Press, 1984, 55–72.

Jaffe, Adam, "Technological Opportunity and Spillovers of R&D," *American Economic Review*, 1986, *76(5)*, 984–1001.

Jorgenson, Dale W., "Microeconomics and Productivity." In Landau, R., and N. Rosenberg, eds., *The Positive Sum Strategy*. Washington, D.C.: National Academy of Sciences, 1986, 57–76.

Jorgenson, Dale, F. Gollop, and B. Fraumeni, *Productivity and U.S. Economic Growth*, Cambridge, MA.: Harvard University Press, 1988.

Kendrick, J. W., "International Comparisons of Recent Productivity Trends." In Schurr, S. H., ed., *Energy, Productivity, and Economic Growth*. Cambridge, MA.: Oelgeschlager Gunn and Hain, 1983, 71–120.

Lichtenberg, Frank, "The Relationship Between Federal Contract R&D and Company R&D," *American Economic Association Papers and Proceedings*, May 1984, *74(2)*, 73–78.

Link, A. N., *Research and Development Activity in U.S. Manufacturing*. New York: Praeger, 1981a.

Link, A. N., "Basic Research and Productivity Increase in Manufacturing: Additional Evidence," *American Economic Review*, 1981b, *71*, 1111–1112.

Maddison, Angus, "Growth and Slowdown in Advanced Capitalist Economies," *Journal of Economic Literature*, 1987, *25(2)*, 649 698.

Mansfield, Edwin, "Rates of Return from Industrial Research and Development," *American Economic Review*, 1965, *LV(2)*, 310–322.

Mansfield, Edwin, "Basic Research and Productivity Increase in Manufacturing," *American Economic Review*, December 1980, *70(5)*, 863–873.

Mansfield, Edwin et al., "Social and Private Rates of Returns from Industrial Innovations," *Quarterly Journal of Economics*, 1977, *91(2)*, 221–240.

National Science Foundation, *National Patterns of Science and Technology Resources 1986*, NSF 86-309, and earlier issues.

Scherer, Fredric M., "Interindustry Technology Flows and Productivity Growth," *Review of Economics and Statistics*, 1982, *64*, 627–634.

Scherer, Fredric M., "Using Linked Patent and R&D Data to Measure Interindustry Technology Flows." In Griliches, Z., ed., *R&D, Patents, and Productivity.* Chicago: University of Chicago Press, 1984, pp. 417–464.

Schmookler, Jacob, *Invention and Economic Growth.* Cambridge, MA.: Harvard University Press, 1966.

Shapiro, Matthew D., "Measuring Market Power in U.S. Industry," National Bureau of Economic Research Working Paper No. 2212, 1987.

Terleckyj, Nestor E., *Effects of R&D on the Productivity Growth of Industries.* Washington, D.C.: National Planning Association Report No. 140, 1974.

U.S. Department of Labor, Bureau of Labor Statistics, "International Comparisons of Manufacturing Productivity and Labor Costs Trends, 1986," Washington: U.S.G.P.O., USDL 87-237.

U.S. Department of Labor, Bureau of Labor Statistics, *Trends in Multifactor Productivity, 1948–81,* Bulletin 2178. Washington, D.C.: U.S.G.P.O., 1983, and subsequent releases.

U.S. Department of Labor, Bureau of Labor Statistics, "Multifactor Productivity Measures, 1986" Washington: U.S.G.P.O., USDL 87-436.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 23–41

Productivity and Postwar U.S. Economic Growth

Dale W. Jorgenson

The purpose of this paper is to analyze the sources of postwar U.S. economic growth. The analysis of sources of economic growth originated in a notable but neglected article by Tinbergen (1942).[1] Among the many remarkable features of Tinbergen's study was an international comparison of growth of output, primary factor input, and total factor productivity for France, Germany, the United Kingdom, and the United States for the period 1870–1914. For the United States, Tinbergen found that the growth of output averaged 4.1 percent per year, the growth of input was 3.0 percent, while productivity growth averaged 1.1 percent. Productivity accounted for only 27 percent of U.S. economic growth during the period 1870–1914.

The findings presented here allocate more than three-fourths of U.S. economic growth during the period 1948–1979 to growth of capital and labor inputs and less than one-fourth to productivity growth. These findings coincide with those of Christensen and Jorgenson (1973) for the period 1929–1969 and the much earlier findings of Tinbergen (1942) for the period 1870–1914. Perhaps surprisingly, many studies of aggregate U.S. economic growth, such as those of Abramovitz (1956) and Solow (1957), have placed greater emphasis on productivity growth than mobilization of capital and labor resources. This is in spite of the fact that these studies have generally allocated about half of economic growth to resource mobilization and half to productivity growth.

To provide additional insight into the sources of U.S. economic growth, this paper then analyzes the sources of growth for individual industrial sectors. The rate of output growth for individual industries is decomposed between the contributions of

---

[1] An English translation appeared in Tinbergen's Selected Papers (1959).

■ *Dale W. Jorgenson is Frederic Eaton Abbe Professor of Economics, Harvard University,*

capital, labor, and intermediate inputs and the growth of sectoral productivity. We find that the contributions of the three inputs account for an overwhelming proportion of the growth of industry output. The final step in this analysis of sources of economic growth is to integrate the sources of output growth for individual industries with sources of growth for the economy as a whole. This enables us to decompose the rate of aggregate productivity growth into a weighted sum of sectoral productivity growth rates and the contributions of reallocations of output and inputs among sectors.

Denison (1985) has analyzed the slowdown in U.S. economic growth since 1973, using an aggregate model of production. He has attributed the slowdown primarily to the decline in aggregate productivity growth. The results presented here bear out his conclusion. The decline in the rate of aggregate productivity growth accounts for 80 percent of the decline in the rate of growth of output. In addition, we find that the slowdown in productivity growth for the economy as a whole is fully explained by the decline in productivity growth at the sectoral level.

However, the decline in economic growth would be left unexplained without an econometric model to determine the rate of productivity growth. Thus, the final objective of this paper is to complete the explanation of the slowdown in U.S. economic growth that took place after 1973. For this purpose we examine econometric models for individual industrial sectors that make the rate of productivity growth for each sector into an endogenous variable. In addition, these models incorporate inputs of energy and materials along with inputs of capital and labor. The models show that higher energy prices are important in explaining the slowdown in U.S. economic growth. An aggregate model of production would fail to establish the critical role of the increase in energy prices after 1973, since energy is excluded as an input at the aggregate level by assumption.

Sources of U.S. Economic Growth

Table 1 allocates aggregate output growth between growth in capital and labor inputs and changes in productivity. The growth rate of output is the sum of the contributions of capital and labor inputs and the rate of productivity growth. The contribution of each input is the product of the value share of the input and its growth rate. The first panel of Table 1 compares growth of output with the contributions of the two primary factor inputs and with productivity growth for the period 1948–1979 and for seven subperiods—1948–1953, 1953–1957, 1957–1960, 1960–1966, 1966–1969, 1969–1973, and 1973–1979.

The endpoints of the periods identified in Table 1 are years in which a cyclical peak occurred. The growth rate presented for each period is the average annual growth rate between cyclical peaks. Capital and labor inputs combined contributed 2.6 percent per year to the output growth rate of 3.4 percent from 1948 to 1979. These two inputs accounted for more than three-fourths of output growth. By contrast, productivity advances contributed only 0.8 percent per year, or only 24 percent of output growth, during this period. The contribution of capital and labor inputs is the

Table 1

Growth in output and its sources, 1948–1979

(measured by average annual rates of growth)

	1948–1979	*1948–1953*	*1953–1957*	*1957–1960*	*1960–1966*	*1966–19698*	*1969–1973*	*1973–1979*
Value added	.0342	.0404	.0302	.0274	.0444	.0326	.0308	.0283
Contribution of capital input	.0156	.0195	.0153	.0105	.0145	.0190	.0156	.0144
Contribution of labor input	.0105	.0124	.0035	.0084	.0141	.0134	.0095	.0105

Source: Jorgenson, Gollop, and Fraumeni (1987).

predominant source of U.S. economic growth for the period as a whole and all seven subperiods.

Comparing the contribution of capital input with other sources of output growth for the period 1948–1979 as a whole makes clear that capital input is the most significant source of growth. The contribution of capital input is also the most important source of growth for six of the seven subperiods, while productivity growth is the most important source for only one, 1960–1966. The contribution of capital input exceeds the contribution of labor input for all seven subperiods, while the contribution of labor input exceeds productivity growth in four of the seven subperiods.

In 1979, the output of the civilian economy stood at almost three times the level of output in 1948. The overall conclusion from this evidence is that the driving force behind the expansion of the U.S. economy between 1948 and 1979 has been the growth in capital and labor inputs. Growth in capital input is the most important source of growth in output, growth in labor input is the next most important source, and productivity growth is least important. Clearly, this perspective focuses attention on the mobilization of capital and labor resources rather than advances in productivity.

Table 1 allocates the sources of U.S. economic growth among contributions of growth in capital and labor inputs, and changes in productivity at the aggregate level. However, a major accomplishment of recent research on the sources of U.S. economic growth is the integration of the growth of intermediate, capital, and labor inputs at the level of individual industrial sectors into the analysis of the sources of growth for the economy as a whole. This makes it possible to attribute U.S. economic growth to its sources at the level of individual industries.

The analysis of sources of growth at the industry level is based on the decomposition of the growth rate of sectoral output into the sum of the contributions of intermediate, capital and labor inputs and the growth of sectoral productivity. The contribution of each input is the product of the value share of the input and its growth rate. Table 2 compares the growth rate of output with the contributions of the three inputs and the growth of productivity for the period 1948–1979. The sum of the

Table 2

Growth in sectoral output and its sources, 1948–1979

(average annual rates of growth)

		Contributions to growth in output			
Industry	*Output*	*Intermediate Input*	*Capital Input*	*Labor Input*	*Rate of Productivity Growth*
Agricultural Production	0.0216	0.0128	0.0040	-.0097	0.0146
Agricultural Services	0.0297	0.0173	0.0068	0.0111	-.0055
Metal Mining	0.0142	0.0075	0.0127	-.0004	-.0056
Coal Mining	0.0038	0.0112	0.0069	-.0116	-.0027
Crude Petroleum and Natural Gas	0.0213	0.0291	0.0104	0.0031	-.0214
Nonmetallic Mining and Quarrying	0.0409	0.0153	0.0178	0.0039	0.0038
Contract Construction	0.0271	0.0169	0.0022	0.0074	0.0006
Food and Kindred Products	0.0281	0.0134	0.0018	-.0002	0.0131
Tobacco Manufactures	0.0072	0.0116	0.0039	-.0015	-.0068
Textile Mill Products	0.0345	0.0157	0.0026	-.0030	0.0192
Apparel and Other Fabricated Textile Products	0.0264	0.0130	0.0016	0.0009	0.0109
Paper and Allied Products	0.0401	0.0322	0.0054	0.0041	-.0016
Printing and Publishing	0.0323	0.0176	0.0023	0.0066	0.0058
Chemicals and Allied Products	0.0591	0.0329	0.0091	0.0050	0.0121
Petroleum and Coal Products	0.0271	0.0422	0.0024	0.0005	- .0179
Rubber and Misc. Plastic Products	0.0477	0.0259	0.0062	0.0113	0.0043
Leather and Leather Products	-.0047	-.0023	0.0005	-.0054	0.0025
Lumber and Wood Products Except Furniture	0.0288	0.0245	0.0045	-.0011	0.0009
Furniture and Fixtures	0.0373	0.0281	0.0029	0.0038	0.0026
Stone, Clay and Glass Products	0.0382	0.0281	0.0057	0.0038	0.0007
Primary Metal Industries	0.0128	0.0154	0.0023	0.011	-.0059
Fabricated Metal Industries	0.0350	0.0200	0.0037	0.0062	0.0050
Machinery except Electrical	0.0417	0.0240	0.0062	0.0080	0.0036
Elec. Machinery					

Table 2
Continued

		Contributions to growth in output			
Industry	*Output*	*Intermediate Input*	*Capital Input*	*Labor Input*	*Rate of Productivity Growth*
Trans. Equipment Except Motor Vehicles	0.0559	0.0408	0.0013	0.0100	0.0039
Motor Vehicles and Equipment	0.0451	0.0285	0.0053	0.0022	0.0091
Prof. Photographic Eqpt. and Watches	0.0569	0.0217	0.0102	0.0170	0.0081
Misc. Manufacturing Industries	0.0340	0.0243	0.0037	0.0015	0.0046
Railroads and Rail Express Service	0.0053	−.0046	0.0019	−.0108	0.0187
Street Rails, Bus Lines and Taxicabs	−.0217	−.0063	0.0036	−.0044	−.0147
Trucking Services and Warehousing	0.0488	0.0222	0.0072	0.0078	0.0116
Water Transportation	0.0040	0.0058	0.0011	−.0019	−.0009
Air Transportation	0.0957	0.0421	0.0103	0.0153	0.0281
Pipelines Except Natural Gas	0.0493	0.0133	0.0282	−.0014	0.0093
Transportation Services	0.0268	0.0488	0.0034	0.0049	−.0304
Tel. and Tel. and Misc. Communication Services	0.0688	0.0077	0.0234	0.0087	0.0290
Radio Broadcasting and Television	0.0132	−.0064	0.0185	0.0216	−.0205
Electric Utilities	0.0628	0.0227	0.0275	0.0034	0.0092
Gas Utilities	0.0531	0.0403	0.0106	0.0024	−.0001
Water Supply and Sanitary Services	0.0328	−.0031	−.0001	0.0048	0.0312
Wholesale Trade	0.0425	0.0064	0.0090	0.0127	0.0145
Retail Trade	0.0293	0.0091	0.0043	0.0056	0.0103
Finance, Insurance and Real Estate	0.0493	0.0341	0.0031	0.0076	0.0044
Services Excluding Private Households and Institutions	0.0377	0.0286	0.0064	0.0078	−.0052
Private Households	0.0491		0.0499	−.0008	
Institutions	0.0373	0.0146	0.0105	0.0182	−.0059
Federal Public Administration	0.0173			0.0173	
Federal Government Enterprises	0.0141			0.0141	
State and Local Educational Services	0.0457			0.0457	
State and Local Public Admin.	0.0362			0.0362	
State and Local Enterprises	0.0363			0.0363	

Source: Jorgenson, Gollop., and Fraumeni (1987).

contributions of intermediate, capital, and labor inputs is the predominant source of growth of output for 46 of the 51 industries included in Table 2.

Comparing the contribution of intermediate input with other sources of output growth demonstrates that this input is by far the most significant source of growth. The contribution of intermediate input exceeds productivity growth and the contributions of capital and labor inputs. If we focus attention on the contributions of capital and labor inputs alone, excluding intermediate input from consideration, these two inputs are a more important source of growth than changes in productivity.

The perspective on U.S. economic growth suggested by the results presented in Table 2 emphasizes the contribution of mobilization of resources within individual industries rather than productivity growth. The explanatory power of this perspective is overwhelming at the sectoral level. For 46 of the 51 industrial sectors included in Table 2, the contribution of intermediate, capital, and labor inputs is the predominant source of output growth. Changes in productivity account for the major portion of output growth in only five industries.

An important feature of productivity growth at the sectoral level is that rates of productivity growth are negative over the period 1948–1979 for a number of industries. These negative growth rates represent a decline in the overall efficiency of production. Some of these declines may be the consequence of the drastic changes in relative prices of capital, labor, and intermediate inputs, especially after the energy crisis of 1973. Cost minimizing firms may have conserved energy by reverting to much earlier vintages of technology for energy utilization. Thus, energy conservation to minimize cost could have resulted in a decline in overall efficiency.

Aggregation Over Sectors

The measures of aggregate output, input, and productivity presented in Table 1 are derived by explicit aggregation over the industrial sectors listed in Table 2. However, the analysis of sources of growth for the U.S. economy as a whole presented in Table 1 is based on a much more restrictive methodology for productivity measurement than the analysis of sources of growth for individual industries given in Table 2. By comparing the two sets of results we can assess the validity of the more restrictive approach employed at the aggregate level.

The methodology we have employed for the economy as a whole is based on an aggregate production function. The existence of an aggregate production function implies very stringent limitations on the underlying sectoral models of production. In particular, we must introduce the concept of value added at the sectoral level.[2] The first step is to define value added as a function of capital and labor inputs and the level of productivity. We then express the output of each industrial sector as a function of value added and intermediate input. The separate effects of changes in capital and labor inputs and changes in productivity are transmitted to the level of

[2] The value added approach is utilized by Kendrick (1983).

output through changes in value added. The concept of value added is implicit in the analysis of sources of economic growth for the U.S. economy as a whole presented in Table 1 and all studies at the aggregate level, beginning with Tinbergen (1942).

One can combine value added functions for all industrial sectors with market equilibrium conditions for each factor of production to obtain an aggregate model of production. Aggregate value added is the sum of quantities of value added in all sectors. Market equilibrium conditions take the form of equalities between the supplies of each type of labor and the sums of demands for that type of labor by all sectors. Similarly, market equilibrium implies equalities between the supplies of each type of capital and the sums of demands for that type of capital by all sectors.

The existence of an aggregate production function imposes additional requirements on the underlying sectoral models of production, too.[3] All sectoral value added functions must be identical to the aggregate production function.[4] In addition, the functions giving capital and labor inputs for each sector in terms of their components must be identical to the corresponding functions at the aggregate level. In essence, the value added function and the capital and labor input functions for each sector must be replicas of the aggregate functions.

The measures of value added, capital input, and labor input presented in Table 1 are constructed from unweighted sums of value added, components of capital input, and components of labor input over all industries. An alternative measure of aggregate value added can be constructed by weighting value added in each industry by the price of value added in that industry. Similarly, alternative measures of aggregate capital and labor inputs can be constructed by weighting individual components of these inputs in each industry by the prices of these components in that industry.

Differences between growth rates of measures of output and input that reflect differences in prices among industries and measures that do not reflect these differences are presented in Table 3. Differences in these growth rates can be interpreted as the contributions of reallocations of value added, capital input, and labor input among sectors to aggregate productivity growth. For example, if labor input is reallocated from a sector with high wages to a sector with low wages, the rate of aggregate productivity growth increases with no corresponding increase in the rates of sectoral productivity growth. The rate of aggregate productivity growth can be represented as a weighted sum of sectoral productivity growth rates and the reallocations of value added and capital and labor inputs.

Reallocations of value added incorporate differences in value added functions among industries. The reallocations also incorporate departures from the assumptions required for the existence of a value added function for each industrial sector. Similarly, reallocations of capital and labor inputs incorporate differences in these aggregates among sectors as well as departures from the assumptions required for the existence of these aggregates within each sector. If the prices of value added, capital input, and labor input were the same for all industries or if value added and all

---

[3] The implications of aggregation over industrial sectors for the existence of an aggregate production function was a central issue in the "reswitching controversy" initiated by Samuelson (1962).

[4] This condition for the existence of an aggregate production function is due to Hall (1973)

Table 3
Growth in aggregate productivity and its sources, 1948–1979
(average annual rates of growth)

Variable	*1948–1979*	*1948–1953*	*1953–1957*	*1957–1960*	*1960–1966*	*1966–1969*	*1969–1973*	*1973–1979*
Rate of Productivity Growth	.0081	.0085	.0113	.0084	.0159	.0002	.0058	.0034
Rate fo Sectoral Productivity Growth	.0083	.0177	.147	.0115	.0162	.0013	.0044	−.0072
Reallocation of Value Added	−.0009	−.0088	−.0040	−.0016	−.0022	−.0019	.0014	.0087
Reallocation of Capital Input	.0009	−.0001	.0020	.0015	.0018	.0012	.0009	−.0001
Reallocation of Labor Input	−.0002	−.0002	−.0014	−.0030	.0000	−.0004	−.0009	.0021

Source: Jorgenson, Gollop, and Fraumeni (1987).

capital and labor inputs were to grow at the same rate, all the reallocations would vanish.

The data presented in Table 3 make it possible to assess the significance of departures from the assumptions that underly the aggregate model of production. Over the period 1948–1979 the reallocations are very small relative to the growth of capital and labor inputs and productivity growth. These results show that an aggregate production model may be appropriate for studies of long-term U.S. economic growth. However, an aggregate model can be seriously misleading for relatively short time periods, such as the individual business cycles 1948–1953, 1966–1969, and 1973–1979.[5]

More specifically, the rate of aggregate productivity growth is 0.81 percent per year, while the weighted sum of sectoral productivity growth rates is 0.83 percent per year for the period 1948–1979 as a whole. The difference between the aggregate productivity growth rate and the weighted sum of sectoral productivity growth rates provides a measure of departures from the stringent assumptions that underly the aggregate production model. This difference is equal to the sum of the reallocations of outputs and inputs among sectors. The difference is small, but not negligible, for the period 1948–1979.

Departures from the Aggregate Production Model

To illustrate the limitations of the aggregate production model, consider the slowdown in U.S. economic growth since 1973. For the period 1948–1979 the annual growth rate averages 3.42 percent. During the subperiod 1973–1979 the average growth rate is only 2.83 percent, a decline of 0.59 percent. The contribution of capital

[5] Gollop (1985) has surveyed the literature on the role of intersectoral shifts.

input declined 0.12 percent per year between the two periods, while the contribution of labor input was unchanged at 1.05 percent. By contrast the decline in the productivity growth rate was 0.47 percent, almost four-fifths of the decline in the growth rate of output. We conclude that the most important factor in the slowdown is the decline in productivity growth.

We can decompose the decline in the aggregate productivity growth rate into the weighted sum of sectoral productivity growth rates and the reallocations of value added, capital input, and labor input. The decline in sectoral productivity growth between the period 1948–1979 and the subperiods 1973–1979 was 1.55 percent per year. This decline is more than sufficient to account for the slowdown in U.S. economic growth. The precipitous fall in sectoral productivity growth was masked at the aggregate level by increases in the reallocations of value added and labor input of 0.96 percent and 0.23 percent, respectively. The reallocations of capital input declined by 0.10 percent.

The conclusion should be clear: The assumptions that underly the aggregate model of production fail to hold during the period 1973–1979. Unfortunately, this model has been used by Denison and other analysts as the sole basis for assessing the determinants of the slowdown in U.S. economic growth during this period. Since the stringent assumptions required for the validity of the model fail to hold during the period, this assessment of the causes of the slowdown can be seriously misleading.

Econometric Modeling of Production

The previous section shows that the slowdown in U.S. economic growth since 1973 can be attributed to a decline in aggregate productivity growth. Closer analysis reveals that this decline can be traced, in turn, to slower growth in productivity at the level of individual industries. However, in the absence of an econometric model to explain the rate of productivity growth at the sectoral level, the decline in U.S. economic growth would be left unexplained.

A wide variety of alternative production models are available for both aggregate and sectoral production modeling. The aggregate production model introduced by Cobb and Douglas (1928) and developed by Tinbergen (1942) retains its usefulness in modeling long term growth trends. However, the critical evidence provided by the energy crises of the 1970s has exposed important empirical limitations of aggregate production modeling. These limitations cannot be overcome by introducing additional complexity at the aggregate level. In short, sectoral production models are essential in assimilating the important new evidence on patterns of production made available by the energy crisis of the 1970s and its aftermath.

The benefits of an aggregate production model must be weighed against the costs of departures from the highly restrictive assumptions that underly the existence of an aggregate production function. We have shown that these assumptions are inappropriate during the period 1973–1979. Where an aggregate model of production is inappropriate, an econometric approach based on sectoral models of production can be employed in analyzing patterns of production for the economy as a whole.

Sectoral Production Modeling

I will argue that the decline in productivity growth at the level of individual industries is the main culprit in the slowdown of U.S. economic growth since 1973. Thus, to provide an explanation of this decline requires identifying the determinants of productivity growth at the sectoral level. To illustrate the econometric approach to productivity growth we summarize the results of fitting an econometric model to data on output and inputs for industrial sectors of the U.S. economy.

Index numbers for intermediate, capital, and labor inputs and rates of productivity growth are employed in the analysis of sources of economic growth presented in Table 2. Production functions can be used in specifying econometric models of production. These models determine the distribution of the value of output among the inputs and the rate of productivity growth. In estimating the parameters of the econometric models the quantity indexes of inputs, the corresponding price indexes, and indexes of productivity growth can be employed as data.

The econometric study summarized here is based on sectoral models of production for each of 35 industries. These industries represent a consolidation of the 51 sectors included in Table 2. Although production functions contain all the available information about producer behavior, it is useful here to express the sectoral models of production in an alternative and equivalent form. Under the assumption of constant returns to scale we can introduce price functions for each industry.[6] The price function gives the price of output as a function of the prices of all inputs and of time, the variable that represents the level of technology. These price functions just summarize the available information about producer behavior in a more convenient form.

To explain the cause of the productivity slowdown, it is also useful to disaggregate intermediate inputs between energy and materials inputs. Given the price function for each industry, we express the shares of each of the four inputs in the industry—capital, labor, energy, and materials inputs—in the value of output as functions of the prices of inputs and the level of technology. Finally, the model is completed with an equation that gives the rate of productivity growth as a function of the prices of the inputs and the level of technology. This equation is our econometric model of sectoral productivity growth.

As in any econometric model, the relationships determining the rate of productivity growth and the value shares of capital, labor, energy, and materials inputs involve unknown parameters. Included among these parameters are biases of productivity growth. For example, the bias for capital input gives the change in the value share of capital in response to changes in technology.[7] We say that productivity growth is

---

[6] The price function was introduced by Samuelson (1953).

[7] This definition of the bias of productivity growth is due to Hicks (1963). Alternative definitions of biases of productivity growth are compared by Binswanger (1978).

Table 4

Classification of industries by biases of productivity growth

Pattern of biases	*Industries*
Capital using Labor using Energy using Material saving	Agriculture, metal mining. crude petroleum and natural gas, nonmetalic mining, textiles, apparel, lumber, furniture, printing, leather, fabricated metals, electrical machinery, motor vehicles, istruments, miscellaneous manufacturing, transportation, trade, finance, insurance and real estate, services
Capital using Labor using Energy saving Material saving	Coal mining, tobacco manufactures, communications, government enterprises
Capital using Labor saving Energy using Material saving	Petroleum refining
Capital using Labor saving Energy saving Material using	Construction
Capital saving Labor saving Energy using Material saving	Electric utilities
Capital saving Labor using Energy saving Material saving	Primary metals
Capital saving Labor using Energy using Material saving	Paper, chemicals, rubber, stone, clay and glass, machinery except electrical, transportation equipment and ordnance, gas utilities
Capital saving Labor saving Energy using Material using	Food

Source: Jorgenson and Fraumeni (1981), p. 44.

capital using if the bias for capital input is positive. Similarly, we say that productivity growth is *capital saving* if the bias for capital input is negative. The sum of the biases for all four inputs must be precisely zero, since the changes in all four value shares must sum to zero.

The biases of productivity growth appear as coefficients of time, representing the level of technology, in the equations for the value shares of the four inputs. The biases appear with an opposite sign as coefficients of the prices in the equation for productivity growth. This feature makes it possible to use information about changes in productivity growth with prices and changes in value shares with technology in estimating the biases of productivity growth as unknown parameters of an econometric model.[8]

Capital using productivity growth implies that an increase in the price of capital input diminishes the rate of productivity growth. Similarly, capital saving productivity growth implies that productivity growth increases with the price of capital input. Jorgenson and Fraumeni (1981) have fitted econometric models based on translog price functions to data for the aforementioned 35 industrial sectors. Since our primary concern is with the determinants of sectoral productivity growth, Table 4 presents a classification of industries by biases of productivity growth.

The pattern of productivity growth that occurs most frequently in Table 4 is capital using, labor using, energy using, and materials saving productivity growth. This pattern occurs for 19 of the 35 industries. For this pattern the biases of productivity growth for capital, labor, and energy inputs are positive and the bias of productivity growth for materials input is negative. This pattern implies that increases in the prices of capital, labor and enrgy inputs diminish the rate of productivity growth, while an increase in the price of materials input enhances productivity growth.

The Slowdown in Productivity Growth

The most striking change in the relative prices of capital, labor, energy, and materials inputs that has taken place since 1973 is the substantial increase in the price of energy. Reversing historical trends toward lower real prices of energy in the United States, the Arab oil embargo of late 1973 and early 1974 resulted in a dramatic increase in oil import prices. Real energy prices to final users increased by 23 percent in the United States during the period 1973–1975, despite price controls on domestic petroleum and natural gas. In 1978 the Iranian revolution sent a second wave of oil import price increases through the U.S. economy. Real energy prices climbed by 34 percent over the following two years.[9]

This evidence provides part of the solution of the problem of disappointing U.S. economic growth since 1973. Higher energy prices are associated with a decline in sectoral productivity growth for 29 of the 35 industries included in Table 4. The resulting slowdown in sectoral productivity growth is more than sufficient to explain the decline in U.S. economic growth. The results of Jorgenson and Fraumeni (1981) have been corroborated by Jorgenson (1984) by further disaggregating energy between electricity and nonelectrical energy.

It is important to emphasize once again that an econometric model of sectoral productivity growth is essential to solving the problem of the slowdown in U.S.

---

[8] Further details on econometric modeling of sectoral productivity growth are given by Jorgenson (1986a).

[9] Trends in energy prices since 1973 are discussed in greater detail by Jorgenson (1986b). Bruno (1984) has discussed the impact of higher raw materials prices after 1973. However, the bias of productivity growth for materials is positive for 33 of the 35 industries listed in Table 4. For these industries an increase in the price of materials is associated with higher rather than lower productivity growth.

economic growth since 1973. An aggregate model of production excludes energy and materials inputs by definition, since deliveries to intermediate demand are offset by receipts of intermediate inputs.

Conclusion

The analysis of sources of economic growth is an essential component of any study of economic growth. However, a theory of growth must also include an explanation of the growth in supplies of capital and labor inputs. In the neoclassical model of economic growth presented by Tinbergen (1942), saving generates growth in capital input and population growth generates growth in labor input. These features of the theory of economic growth have been retained in the neoclassical growth models developed by Solow (1956) and Tobin (1955).

The theoretical underpinnings of an analysis of growth in factor supplies are to be found in the theory of consumer behavior. For example, the study of saving requires modeling saving-consumption decisions. Similarly, the analysis of labor supplies requires modeling demographic behavior and labor-leisure choices. A theory of economic growth must incorporate the sources of economic growth and the modeling of producer behavior. The analysis of growth of factor supplies and the modeling of consumer behavior are required to complete the theory. Future research should give additional attention to growth of factor supplies and the modeling of consumer behavior. This focus characterized the classic studies of economic growth by Kuznets (1971) and Schultz (1961).

Recent research on economic growth has emphasized the sources of economic growth and the modeling of producer behavior. This has proved to be very fruitful, as suggested by the research summarized above. The steps outlined here—disaggregating the sources of economic growth to the sectoral level, decomposing sectoral output growth between productivity growth and the growth of capital, labor, energy, and materials inputs, and modeling the rate of growth of sectoral productivity growth rate econometrically—have been taken only recently. Much additional research will be required to provide an exhaustive explanation of the slowdown of U.S. economic growth and the implications for the future growth of the economy.

Appendix
Methodology and Data Sources

An excellent overview of research on sources of economic growth, including alternative data sources and methodologies, is provided by the Rees Report to the National Research Council (1979). Tinbergen's methodology has dominated subsequent research on the sources of economic growth. This methodology was based on an aggregate production function for each country. Capital and labor inputs were weighted by their marginal products in determining an index of productivity ("efficiency" in Tinbergen's terminology).[10] Christensen, Cummings, and Jorgenson (1980) and Maddison (1987) have reviewed international comparisons of sources of economic growth among industrialized countries.

The methodology that underlies the data presented in Table 2 is based on sectoral production functions of the translog form introduced by Christensen, Jorgenson, and Lau (1973).[11] Given translog production functions for all sectors, the corresponding price and quantity index numbers can be generated for all three inputs. The growth rate of each input is a weighted average of growth rates of its components. Similarly, the translog index of productivity growth is the difference between the growth rate of output and a weighted average of growth rates of intermediate, capital and labor inputs.[12]

The critical innovation in the methodology that underlies Table 2 is to distinguish among components of intermediate, capital, and labor inputs that differ in marginal productivity. For each sector, intermediate input is represented as a function of deliveries from all other sectors. Capital input is broken down by class of asset and legal form of organization. Finally, labor input is broken down by characteristics of individual workers such as sex, age, education, employment status, and occupation.

Measuring Labor Input

A novel feature of the quantity indexes of labor input presented in Table 2 is that these indexes incorporate data from both establishment and household surveys. Estimates of employment, hours worked, and labor compensation for each industrial sector are controlled to totals based on establishment surveys that underly the U.S. national income accounts. These totals are allocated among categories of the work force cross-classified by the characteristics of individual workers on the basis of household surveys. The resulting estimates of hours worked and average compensation per hour for each sector provide the basis for the indexes of labor input presented in Table 2.

For each of 51 industrial sectors listed in Table 2 prices and quantities of labor input are cross-classified by the two sexes, eight age groups, five educational groups, two employment classes, and ten occupational groups. Annual data from 1948 to 1979 on hours worked and average labor compensation per hour are required for 1600 components of the work force in each industry. For this purpose employment, hours, weeks, and labor compensation within each sector are allocated on the basis of the available cross-classifications. This methodology makes it possible to exploit all the published detail on labor input from the decennial Census of Population and the Current Population Survey.

[10] The concept of total factor productivity was developed independently of Tinbergen's work by Stigler (1947).

[11] Detailed references to sectoral production studies incorporating intermediate input are given by Jorgenson (1986a).

[12] Translog quantity indexes were employed by Christensen and Jorgenson (1969).

Quantity indexes of labor input are presented in Table 2 for 51 industrial sectors. The data on labor input that underly Table 2 are cross-classified by sex, age, education, class of employment, occupation, and industry for a total of 81,600 types of labor input. The growth of labor input can be decomposed to obtain the contributions of all six industrial, occupational, and demographic characteristics.[13]

Measuring Capital Input

The approach to the construction of the data on capital input presented in Table 2 is strictly analogous to the approach for data on labor input. Capital services represent the quantity of capital input, just as labor services represent the quantity of labor input. Measures of capital services for depreciable assets are derived by representing capital stock at each point of time as a weighted sum of past investments. The weights correspond to the relative efficiencies of capital goods of different ages, so that all components of capital stock have the same efficiency when appropriately weighted.

Rental rates for capital services provide the basis for property compensation, just as wage rates provide the basis for labor compensation. Information on rental transactions would be required to employ data sources for capital input that are analogous to those we have used for labor input. These data are not available in a readily accessible form, even for the substantial proportion of assets with active rental markets. However, rental values can be imputed on the basis of estimates of capital stocks and property compensation.

Data on rental prices for depreciable assets are generated by allocating property compensation for return to capital, depreciation, and taxes among assets. Depreciation is the decline in value of a capital good with age at a given point in time, so that estimates of depreciation depend on the relative efficiencies of capital goods of different ages. The estimates of capital input presented in Table 2 incorporate the same data on relative efficiencies of capital goods into estimates of both capital stocks and rental prices.

Data on capital input are unavailable for the five government sectors listed in Table 2. For each of the 45 private industrial sectors listed in this table, capital input is cross-classified by four asset classes (producers' durable equipment, nonresidential structures, inventories, and land) and two legal forms of organization (corporate and noncorporate business). Households and institutions are treated as a separate sector with prices and quantities of capital input cross-classified by producers' and consumers' durable equipment, residential and nonresidential structures, and land.

Data on producers' durable equipment can be further subdivided among 20 categories, while data on nonresidential structures can be subdivided among 14 categories. Finally, data for each category of assets can be distinguished by age. On average, 43.5 vintages are distinguished for each category of producers' durable equipment and 73.5 vintages are distinguished for nonresidential structures. For each of the 45 private industrial sectors listed in Table 2, annual data from 1948 to 1979 on capital stock and its rental price are required for an average of as many as 3802 components of the capital stock.

---

[13] Chinloy (1981) provides such a decomposition for the U.S. economy.

The first step in developing sectoral measures of capital input is to construct estimates of capital stock by industry for each year from 1948 to 1979. Investment data for producers' durable equipment and structures are distributed among industries on an establishment basis. Estimates of investment for all sectors are controlled to totals from the U.S. national product accounts. For residential structures, investment data are taken directly from the U.S. national product accounts. Investment goods prices from the U.S. national product accounts are employed to obtain estimates of investment in equipment and structures in constant prices.

The second step in developing sectoral measures of capital input is to construct estimates of prices of capital services from data on property compensation. For each asset the price of investment goods is a weighted sum of future rental prices, discounted by a factor that incorporates future rates of return. Weights are given by the relative efficiencies of capital goods of different ages. The same weights are used in constructing estimates of rental prices and capital stocks. For depreciable assets the weights decline with age; for nondepreciable assets the weights are constant.

Measuring Output, Intermediate Input, and Productivity

An important innovation embodied in the data on productivity presented in Table 2 is that intermediate, capital, and labor inputs are treated symmetrically at the sectoral level. The value of output at the sectoral level includes the value of intermediate input as well as capital and labor inputs. All three inputs are employed in analyzing the sources of growth in sectoral output. The industry definitions employed in the U.S. national income accounts are used in measuring output. These definitions are based on establishments within each industry.

Data on output in current and constant prices are available from the Interindustry Economics Division of the Bureau of Economic Analysis (BEA, 1974) for the manufacturing sectors and from a study by Jack Faucett Associates (1975) for the Bureau of Labor Statistics (BLS) for nonmanufacturing sectors. To evaluate output from the point of view of the producing sector, excise and sales taxes must be subtracted and subsidies must be added to the value of output. The resulting price of output from the producers' point of view is equal to the ratio of the value of output in current prices to the value of output in constant prices.

Data on interindustry transactions published by (BEA) must be employed to disaggregate intermediate input by sector of origin. These data are based on industry definitions employed in the U.S. interindustry accounts. To bring measures of intermediate input into conformity with industry definitions from the U.S. national income accounts, interindustry transactions must be reallocated among sectors. To construct prices and quantities of intermediate input by sector of origin the value of intermediate input originating in each sector must be deflated by an index of purchasers' prices for the output of that sector.

For the five government sectors included in Table 2 output is equal to labor input by definition; for private households output is equal to an index of capital and labor input. For these six sectors productivity growth is zero by definition. Rates of productivity growth for the remaining 45 sectors are presented on an annual basis for the period 1948–1979 in Table 2.

The findings on the sources of U.S. economic growth summarized in Tables 1 and 2 have been generated by a truly massive empirical research effort. The data on labor input have incorporated all the annual detail on employment, weeks, hours worked, and labor compensation published in the decennial Census of Population and the Current Population Survey. Similarly, the data on capital input have incorporated all the available detail on investment in capital goods by industry and class of asset and on property compensation by legal form of organization from the U.S. national income and product accounts. Finally, the data on intermediate input and output have incorporated all of the available annual data by industry from the U.S. national income and product accounts and the U.S. interindustry accounts.

Aggregation Over Sectors

Aggregate data on prices and quantities of labor input, cross-classified by sex, age, education, employment class, and occupation, but not by industry, underly the indexes of labor input presented in Table 1. For the economy as a whole hours worked and labor compensation for each of 1600 categories of the work force are added over all industries. Labor compensation is divided by annual hours worked to derive labor compensation per hour worked for each category. Finally, price and quantity data are combined into price and quantity indexes of aggregate labor input.

Aggregate data on prices and quantities of capital input, cross-classified by asset class, legal form of organization, and age, but not by industry, underly the indexes of capital input presented in Table 1. For the economy as a whole, capital stock and property compensation for each category are added over all industries. Property compensation is divided by capital stock to derive property compensation per unit of capital stock for each category. Finally, price and quantity data are combined into price and quantity indexes of aggregate capital input.

The methodology we have outlined for the economy as a whole can be implemented by considering specific forms for the aggregate production function and for capital and labor inputs as functions of their components. We take these functions to be translog in form, so that we can generate a translog index of the rate of productivity growth. The average rate of productivity growth is the difference between the growth rate of value added and a weighted average of growth rates of capital and labor inputs. Similarly, we can generate translog indexes of capital and labor inputs, giving the growth rate of each input as a weighted average of growth rates of its components.

At a methodological level, the integration of data generation and econometric modeling is an important achievement of recent research on the sources of economic growth. The extensive data development we have outlined is firmly rooted in the economic theory of production. The conceptual basis for the measures of intermediate, capital, and labor inputs in Tables 1 and 2 is provided by the theory of exact index numbers employed by Diewert (1976). Diewert showed that the index numbers utilized, for example, by Christensen and Jorgenson (1969) could be generated from the translog production function introduced by Christensen, Jorgenson, and Lau (1973).

The application of the theory of index numbers to the measurement of labor input requires weighting the components of labor input by wage rates. This was first carried out for all industrial sectors of the U.S. economy by Gollop and Jorgenson (1980). Similarly, the measurement of capital as a factor of production involves weighting the components of capital input by rental rates. The conceptual basis for imputing rental prices for capital goods was established by Jorgenson (1963). These rental prices were employed in aggregate productivity measurement by Jorgenson and Griliches (1967) and at the sectoral level by Fraumeni and Jorgenson (1986) and Gollop and Jorgenson (1980).

The final step in the methodology for analyzing sources of economic growth is to aggregate over industrial sectors. This step is critical in integrating the sources of growth for individual industries into the sources of growth for the economy as a whole. The methodology for aggregation over sectors originated by Domar (1961) has been generalized by Hulten (1978) and Jorgenson (1980). This methodology was implemented for the U.S. by Fraumeni and Jorgenson (1986) and underlies that data on aggregate productivity change presented in Table 3.

References

Abramovitz, Moses, "Resources and Output Trends in the United States since 1870," *American Economic Review*, May 1956, *46 (2)*, 5–23.

Binswanger, Hans P., "Issues in Modeling Induced Technical Change." In Binswanger, Hans P., and Vernon W. Ruttan, eds., *Induced Innovation*. Baltimore: Johns Hopkins University Press, 1978, pp. 128–163.

Bruno, Michael, "Raw Materials, Profits, and the Productivity Slowdown," *Quarterly Journal of Economics*, February 1984, *99 (1)*, 1–30.

Bureau of Economic Analysis, U.S. Department of Commerce, *Gross National Product by Industry*, Work File 1205-04-06, Washington, 1974.

Bureau of Economic Analysis, U.S. Department of Commerce, *The National Income and Product Accounts of the United States, 1929–1974*; statistical tables, Washington, U.S.G.P.O., 1977.

Chinloy, Peter, *Labor Productivity*. Cambridge: Abt Books, 1981.

Christensen, Laurits R., and Dale W. Jorgenson, "The Measurement of U.S. Real Capital Input, 1929–1967," *Review of Income and Wealth*, December 1969, *15 (4)*, 293–320.

Christensen, Laurits R., and Dale W. Jorgenson, "Measuring the Performance of the Private Sector of the U.S. Economy, 1929–1969." In Moss, Milton, ed., *Measuring Economic and Social Performance*. New York; National Bureau of Economic Research, 1973, pp. 233–238.

Christensen, Laurits R., Dale W. Jorgenson, and Lawrence J. Lau, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, February 1973, *55 (1)*, 28–45.

Christensen, Laurits R., Dianne Cummings, and Dale W. Jorgenson, "Economic Growth, 1947–1973: An International Comparison. "In Kendrick, John W., and Beatrice Vaccara, eds., *New Developments in Productivity Measurement*. Chicago: University of Chicago Press, 1980, pp. 595–698.

Cobb, Charles W., and Paul H. Douglas, "A Theory of Production," *American Economic Review*, March 1928, *18 (1)*, 139-165.

Denison, Edward F., *Trends in American Economic Growth, 1929–1982.* Washington: The Brookings Institution, 1985.

Diewert, W. Erwin, "Exact and Superlative Index Numbers," *Journal of Econometrics,* May 1976, *4 (2),* 115–146.

Domar, Evsey, "On the Measurement of Technological Change," *Economic Journal,* December 1961, *71,* 709–729.

Fraumeni, Barbara M., and Dale W. Jorgenson, "The Role of Capital in U.S. Economic Growth, 1948–1979." In Dogramaci, Ali, ed., *Measurement Issues and Behavior of Productivity Variables.* Boston: Martinus Nijhoff, 1986, pp. 161–244.

Gollop, Frank M., and Dale W. Jorgenson, "U.S. Productivity Growth by Industry, 1947–1973." In Kendrick, John W., and Beatrice Vaccara, eds., *New Developments in Productivity Measurement.* Chicago: University of Chicago Press, 1980, pp. 17–136.

Gollop, Frank M., "Analysis of the Productivity Slowdown: Evidence for a Sector-Biased or Sector-Neutral Industrial Strategy." In Baumol, William J., and Kenneth McLennan, eds., *Productivity Growth and U.S. Competitiveness.* New York: Oxford University Press, 1985, pp. 160–186.

Hall, Robert E., "The Specification of Technology with Several Kinds of Output," *Journal of Political Economy,* July/August 1973, *81 (4),* 878–892.

Hicks, John R., *The Theory of Wages,* 2nd ed. (1st ed. 1932), London, Macmillan, 1963.

Hulten, Charles R., "Growth Accounting with Intermediate Inputs," *Review of Economic Studies,* October 1978, *45 (3),* 511–518.

Jack Faucett Associates, *Output and Employment for Input-Output Sectors,* Washington, D.C.: U.S. Bureau of Labor Statistics, March 1975.

Jorgenson, Dale W., "Capital Theory and Investment Behavior," *American Economic Review,* May 1963, *53 (2),* 247–259.

Jorgenson, Dale W., and Zvi Griliches, "The Explanation of Productivity Change," *Review of Economic Studies,* July 1967, *34 (3),* 249–283.

Jorgenson, Dale W., "Accounting for Capital." In von Furstenberg, George, ed., *Capital, Efficiency and Growth.* Cambridge: Ballinger, 1980, pp. 251–319.

Jorgenson, Dale W., and Barbara M. Fraumeni, "Relative Prices and Technical Change." In Berndt, Ernst R., and Barry C. Field, eds., *Modeling and Measuring Natural Resource Substitution.* Cambridge: M.I.T. Press, 1981, pp. 17–47.

Jorgenson, Dale W., "The Role of Energy in Productivity Growth." In Kendrick, John W., ed., *International Comparisons of Productivity and Causes of the Slowdown.* Cambridge: Ballinger, 1984, pp. 270–323.

Jorgenson, Dale W., "Econometric Methods for Modeling Producer Behavior." In Griliches, Zvi, and Michael D. Intriligator, eds., *Handbook of Econometrics,* vol. 3. Amsterdam: North-Holland, 1986a, pp. 1841–1915.

Jorgenson, Dale W., "The Great Transition: Energy and Economic Change," *Energy Journal,* 1986b, *7 (3),* 1–13.

Jorgenson, Dale W., Frank M. Gollop, and Barbara M. Fraumeni, *Productivity and U.S. Economic Growth,* Cambridge, Harvard University Press, 1987.

Kendrick, John W., *Interindustry Differences in Productivity Growth,* Washington, American Enterprise Institute, 1983.

Kuznets, Simon, *Economic Growth of Nations.* Cambridge: Harvard University Press, 1971.

Maddison, Angus, "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment," *Journal of Economic Literature,* June 1987, *25 (2),* 649–698.

National Research Council, *Measurement and Interpretation of Productivity.* Washington: National Academy of Sciences, 1979.

Samuelson, Paul A., "Prices of Factors and Goods in General Equilibrium," *Review of Economic Studies,* 1953, *21 (1),* 1–20.

Samuelson, Paul A., "Parable and Realism in Capital Theory: The Surrogate Production Function," *Review of Economic Studies,* June 1962, *29 (3),* 193–206.

Schultz, Theodore W., "Investment in Human Capital," *American Economic Review,* March 1961, *41 (1),* 1–17.

Solow, Robert M., "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics,* February 1956, *70 (1),* 65-94.

Solow, Robert M., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics,* August 1957, *39 (3),* 312–320.

Stigler, George J., *Trends in Output and Employment,* New York, National Bureau of Economic Research, 1947.

Tinbergen, Jan, "Zur Theorie der Langfristigen Wirtschaftsentwicklung," *Weltwirtschaftliches Archiv,* 1942, *55 (1),* 511–549.

Tinbergen, Jan, *Selected Papers.* Amsterdam: North-Holland, 1959.

Tobin, James, "A Dynamic Aggregative Model." *Journal of Political Economy,* April 1955, *63 (2),* 103–115.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 43–69

The Productivity Slowdown, The Oil Shocks, and the Real Cycle

Mancur Olson

Five facts stand out about economic performance in the United States since World War II. The first is that, for most of the postwar period, the U.S. has had a slower rate of economic growth than the other major developed democracies except Great Britain, and much lower rates of growth than Germany, Japan, and Italy. The gap between the growth of per capita income of the United States and comparable countries has, however, tended to diminish over time and the United States has lately grown at about the same rate as most other developed countries.

The second outstanding fact is that the United States has fallen far behind in many mass production industries in which it used to lead the world, whereas it is ahead of other nations in many new and especially high-technology industries. In automobiles, steel, tires, household appliances, and earth-moving and farm equipment, for example, U.S. production has fallen dramatically in relation to output in most comparable countries and the country has become a net importer of most such products. By contrast, in newer industries, such as computers, jet aircraft, sophisticated weaponry, and high technology in general, the United States is often dominant both as producer and exporter.

The third outstanding fact is that most of the older heartland of American industry—roughly the area north and east of an arc drawn from Baltimore to St. Louis to Milwaukee—has declined dramatically in comparison with the South and the West. This change in regional fortunes has been due in large part to migration, motivated by differences in economic prospects, from the declining to the advancing

■ *Mancur Olson is Professor of Economics, University of Maryland, College Park, Maryland.*

regions.[1] The first states to experience pronounced economic decline were the industrial states of New England, which became "depressed areas" in the 1950s, but this region has lately gained particularly from new high technology industries and its decline has been halted or interrupted. The decline of the industrial areas near the Great Lakes and in the mid-Atlantic states began later, but has continued relentlessly.

The fourth conspicuous fact is the near cessation of the growth of productivity since the early 1970s in the United States as well as in other developed countries. Productivity growth in the 1950s and 1960s had been above historic levels, so a return to typical rates of productivity growth would not have been surprising, but the rates fell far below normal levels and have (in spite of some improvement in the 1980s) remained unusually low.

The fifth dramatic fact is the shift from the unexpectedly good macroeconomic performance of the 1950s and 1960s to the simultaneous and theoretically anomalous increases in unemployment and inflation rates in the 1970s.

The five outstanding facts about postwar American economic performance need to be explained. Yet until recently there has been no explanation whatever of the first three. Some economists have purported to explain the fourth and the fifth in terms of the oil shocks. This article attempts to show, with distinctive quantitative estimates, that the familiar attempts to trace the productivity slowdown to higher energy prices cannot be right, but that the oil shocks had a significant *indirect* impact on the productivity slowdown. I argue that this roundabout impact becomes evident only when we introduce a fresh perspective on macroeconomics that also helps to explain the anomalous emergence of stagflation in the 1970s. Auspiciously, all five prominent facts are explained within the same parsimonious framework, but the explanation is too general to provide the illumination of important details that is ultimately needed.

The Productivity Slowdown and the Oil Shocks

Before Robert Solow's pathbreaking article (1957) and Edward Denison's pioneering book (1962), it had usually been taken for granted that economic growth was due to capital accumulation and to the augmentation of other productive inputs. Solow's suggestive calculations indicated that perhaps seven-eighths of U.S. growth from 1909 to 1949 was due to technical change and one-eighth to the growth of capital. Then Denison's detailed and careful work, showing that there was a large "residual" growth of productivity that could not be explained by the augmentation of capital and other measurable factors, convinced almost all economists that the advance of knowledge was indeed the leading source of economic growth.

---

[1] The progress of regions within countries with freedom of mobility should accordingly be measured by the growth of total rather than per capita income, even though the latter is a good measure of national economic performance. The former captures the migration toward areas of economic growth, whereas the migration normally tends to reduce the growth of per capita income in advancing areas and increase it in declining areas.

Once the great increase in multifactor productivity came to be recognized, it seemed obvious where it came from. The U.S. was spending unprecedented sums on research and development and there were dramatic examples of productivity-increasing technological advances. Though some specialists were relatively cautious, it quickly became conventional wisdom to attribute the vastly greater amounts of growth than could be accounted for by the increase in the productive factors to the "advance of knowledge" that grew out of research and development.

The cozy conviction that the tractor of technological advance would always pull the trailer of multifactor productivity behind it has been shattered: the two became unhitched in about 1973. Though at times the percentage of the GNP devoted to R&D fell a little and federally financed R&D spending in some big manufacturing firms decreased (Griliches, 1986), total public and private expenditures on R&D in constant dollars typically increased: they were over three times as large in 1973 as they had been in 1953, and two-thirds larger than in 1960.[2] In 1980–85, the average level of real expenditures was a fourth higher than in the last five years of the 1960s (National Science Foundation, 1986). Thus reduced spending on R&D cannot explain the productivity slowdown.

There is the possibility that science hit a stone wall in the early 1970s and that R&D expenditures produced no new knowledge. But this possibility is hard to take seriously, in part because almost everyone has observed many dazzling technological discoveries. As Solow (1987) pithily puts it, we have been going through "what everyone feels to have been a technological revolution," and we "see the computer age everywhere but in the productivity statistics."

Unfortunately, when all of the other usual suspects that might be guilty of causing a productivity slowdown are rounded up, they do not, even in the aggregate, explain the drop in productivity growth (Denison, 1979). Many oft-named suspects, such as the growth of the welfare state or a decline in the work ethic, do not have the required time-profile, since the growth of government and any changes in culture have occurred gradually over decades, whereas there was a sudden drop in productivity growth in 1973.

This objection does not, however, apply to one suspect: the sudden increase in the prices of imported oil at the end of 1973. There is the problem that, on closer analysis, the productivity slowdown appears to have started *before* the first oil shock. If the Bureau of Labor Statistics quarterly data on productivity (1974) are right, the productivity slowdown began two quarters before the oil shock, and many studies point to a decline in productivity starting as early as the mid-1960s. Still, all agree that there was a dramatic drop in productivity growth in 1973, so the timing is not far off.

---

[2] Absolute price deflated figures are probably a better measure of the impact of R&D expenditures than percentages: the increased use of an advance in knowledge to produce a larger output does not depreciate or reduce the stock of knowledge.

Though inflation was accelerating well before 1973, the oil shock also has approximately the right initial timing to explain the macroeconomic anomaly of simultaneous inflation and unemployment. The similar timing of the energy crisis, on the one hand, and of both stagflation and the productivity slowdown, on the other, no doubt helps explain why higher oil prices are so often blamed. The oil shocks have been used to justify Keynesian theory in the face of the anomaly of simultaneous inflation and unemployment and, paradoxically, have also helped to justify equilibrium "real business cycle" theories which, in contradiction to traditional Keynesian theories, postulate that all markets clear. This article will later offer a macroeconomic approach that integrates real shocks with an argument that provides a microeconomic basis for some markets that fail to clear.

It is necessary first to estimate the impact of the substitution away from imported oil on the productivity slowdown. Many studies attribute the productivity slowdown to energy prices, yet other economists object that the tail cannot wag the dog—that energy costs are far too small to have such an impact. To help resolve the controversy and explain the fourth outstanding fact, this article will present a new estimate of the *direct* impact of the oil shock on U.S. productivity.

The first step is to take account of the way in which the Bureau of Economic Analysis calculates changes in *real* GNP. The BEA treats an increase in the price of imports in essentially the way it treats domestic prices that rise because of inflation. As Denison (1981) has explained, it deflates the current-dollar prices of U.S. imports by an index of the prices of imports and deflates the current-dollar prices of exports by an index of the prices of exports. One deflated figure is then subtracted from the other to obtain the real net foreign investment total for the calculation of real GNP. This means that when higher prices of imported oil shift the terms of trade against the United States, the conventional measures of real output are *not directly affected*. Obviously the standard of living of Americans is affected, but that is another matter. For studies of the productivity of the U.S. economy, it is appropriate to focus on what the U.S. economy produces and not on international price changes. This is what the official statistics do, and it is in these statistics that the productivity slowdown is found.

Since the higher prices Americans paid for imported oil after 1973 did not enter into the real GNP figures, how did they affect productivity? Whenever resources were diverted from other uses to substitute for more expensive imported oil, the U.S. economy lost whatever output the reallocated resources would have produced in their alternative uses. The U.S. economy obtained some output and some reduction in oil imports from the reallocated resources and if the reallocation was rational it made Americans better off. But the reallocated resources produced less output, at the old prices that are relevant for a calculation of real growth or productivity (Lindbeck, 1983), than they produced before. If the allocation of resources was efficient before the increase in the price of imported oil, any adaptation to the higher price of imported oil had to reduce U.S. productivity.

This concern for the way real GNP is defined leads immediately to two points that are crucial to the analysis here. First, the direct impact of the higher prices of imported oil on the productivity slowdown that is debated in the growth-accounting and econometrics literature could occur only to the extent that the U.S. actually *did substitute for imported oil*: as long as it maintained oil imports, the higher prices could have no direct impact on the productivity slowdown. Second, the total social cost of the higher price of imported oil *greatly exceeded the costs of the substitutions away from imported oil* that directly reduced productivity, because the country had to pay more for the oil it continued to import.

Both types of costs are limited by one factor that is evident the moment we consider a firm in equilibrium. In the absence of indivisibilities, a firm gets the same increment in revenue out of the last dollar spent on each type of input, so it can usually substitute away from any one input to some extent at little cost. Any additional costs will somewhat raise prices, but buyers in equilibrium also get the same gain from the last dollar spent on each good, so they will also usually be able to forego purchasing marginal units of a good that has risen in price without much sacrifice. Since there are characteristically many stages in the economic process and substitution at each stage reduces the cost increase for the next stage, a multilayered set of substitutions usually makes the total losses from a shortage surprisingly small.

Thus relative prices change all the time but normally have no impact on the economy as a whole. Even in cases where shortages are unusually severe because of wartime embargoes, naval blockades, or strategic bombing, the process of substitution normally makes the losses from the disruption of supplies incomparably smaller than expected. The allied bombing campaign against the Nazis, for example, succeeded in the air, but German munitions output *tripled* between 1941 and 1944 (Klein, 1959). The history of strategic bombing, naval blockades, embargoes, and boycotts is overwhelmingly a story of startling failures (Knorr, 1975; Olson, 1962, 1963).

When a shortage is due to a higher world price rather than some wartime limit on availabilities, the higher price sets an upper bound on the costs of substitution. Since it was always possible for the U.S. to buy more oil, obviously no rational substitution by a firm or individual to save a gallon of oil cost could have cost more than the price of a gallon of oil.[3]

A moment's reflection about substitution also tells us what time profile the costs of the substitution away from imported oil must have had. Though some adaptations to higher priced oil could be undertaken quickly, others had to wait until existing capital was replaced. Thus the substitution costs of the higher oil prices must have been sharply limited during the initial years of the oil crisis, but could have become much greater by the 1980s.

Measuring Substitution Costs Correctly is Simpler

Many studies make the mistake of focusing upon the higher price of *energy* in general rather than of imported oil. One reason this is wrong is that the substitution of domestically produced energy for imported oil is part of the cost of the substitution process. Another reason is that it is the impact of an exogenous change in the price of foreign oil that is being debated, so it is the huge rise in the price of oil—not the smaller rise in the price of energy in general—that is at issue. By the same token, it is the substitution away from imported oil, not from all forms of energy, that is pertinent. We are double counting if we treat the higher prices of other forms of domestically produced energy as though they were an *additional* problem rather than a cost-reducing adaptation to higher prices of imported oil.[4] This double-counting is particularly serious for coal, as the United States was a major exporter of coal and profited handsomely from the higher prices of coal that resulted from the higher prices of imported oil.

---

[3] If a country has some monopsony power, as the United States had, the social cost of another barrel of oil includes the increase in the price that the purchase of another barrel would have brought about.

Many estimates of the contribution of the energy crisis to the productivity slowdown also err in considering only the loss on the demand side. As we shall see, the largest substitution of resources arose from the shift of resources into domestic petroleum extraction.

When we properly focus on oil alone and remember the substitutions on the supply side, the main information needed to assess the impact of the oil shortage on the productivity slowdown can be specified in simple supply and demand curves. A profit-seeking producer would not shift resources into domestic oil production unless the marginal cost of the oil produced was expected to be less than or equal to the price, so the value of the extra resources devoted to producing domestic oil is given by a triangle under the supply curve. To the extent that users of oil find it more advantageous to substitute for oil than to pay the higher price, they will buy less, and the measure of their loss is a triangle under the demand curve.

Though domestic price controls made the United States import more oil and raised total costs, they reduced the extent of productivity-reducing substitution. As Figure 1 shows, the relevant triangles are those below the domestic prices. The substitutions on the demand side are given by triangle *gfh* below the new price to domestic buyers, P_b. Though the price to domestic producers depended upon whether the oil was from "old," "new," or stripper wells, it was below the world price during all the early years of the oil crisis. There are many other complications and a different situation in each year, but we can nonetheless illustrate the essence of the situation by assuming a weighted average price to domestic producers, P_p, which leaves the area *abc* as the value of the extra resources devoted to domestic oil production.

The increase in the import bill illustrated by the rectangle *bdeg* can be calculated easily. The rise in the price of oil began at the end of 1973 so we begin by considering 1974. In that year, the United States imported 2149.9 million barrels of oil. The price

---

[4] If other forms of energy were, like domestically produced oil, essentially always perfect substitutes for imported oil, then it would be appropriate to look at energy in general rather than oil. The demand curves even of imperfect substitutes can, however, be interdependent: to the extent that other forms of energy are such important substitutes for oil that modest changes in prices generate big shifts in the demand curve for oil, then these shifts have to be kept in mind in the following analysis.

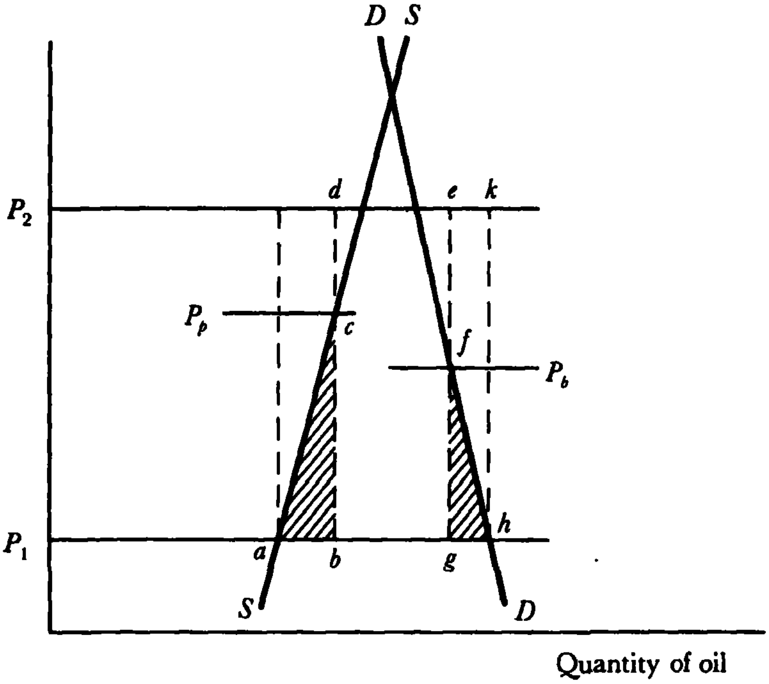


Fig. 1. Illustrating the substitution costs of higher oil prices.

of mideast light crude (in the constant 1982 dollars I shall use throughout this account) was $20.33 a barrel. In 1971-72 the average price was $3.86, virtually the same as it had been over the low-priced decade ending in 1972, so the price had increased by $16.47 in constant dollars. Thus the increase in cost of the 1974 level of imports was simply the 2149.9 million barrels of oil imported times the price increase of $16.47, or $35.41 billion. This was 1.3 percent of 1974 GDP measured in the same 1982 dollars. Table 1 gives the extra cost of the oil actually imported from 1974 on, and how much this cost increased or decreased over each year. In 1979, the year when the bill for the increase in the cost of the oil actually imported was greatest, the bill for oil imports rose by 1.86 percent of that year's GDP. In the peak year of 1979 the oil import bill took 3.17 percent more of that year's GDP to pay than would have been required had the 1971-72 price been paid for the actual 1979 imports. Thus the extra costs of the oil actually imported was large enough so that, *in combination with other factors*, it could have had a significant impact on the U.S. economy.

The *substitution* effects of the oil price increases on the productivity slowdown—the effects that the disputes in the literature are about—are a totally different story. When the matter is looked at the way it is here, the story is surprisingly simple; the approximate cost of the substitutions in the supply side, for example, can simply be looked up in the Census Bureau's *Annual Survey of Oil and Gas*. This source even

Table 1
Increases in Import Bill

				Extra Import Costs		*Annual Increase in Import Bill due to Price Increases*	
Year	*Crude Oil Price in 1982 $ per barrel*	*Price Increase from 1971–72 Average ($3.86)*	*Net Import Quantity (in million barrels)*	*in billions of 1982 dollars*	*as percentage of GDP*	*in billions of 1982 dollars*	*as percentage of GDP*
	(A)	(B)	(C)	(D)	(E)	(F)	(G)
1970	2.88		1153.4				
1971	3.81		1350.5				
1972	3.91		1654.3				
1973	5.68	1.82	2197.3	3.99	0.15	—	—
1974	20.33	16.47	2149.9	35.41	1.31	31.42	1.17
1975	18.06	14.20	2135.3	30.32	1.14	−5.09	−0.19
1976	18.43	14.57	2594.9	37.81	1.35	7.49	0.27
1977	18.68	14.82	3124.4	46.30	1.58	8.48	0.29
1978	17.88	14.02	2920.0	40.94	1.33	−5.35	−0.17
1979	37.94	34.08	2916.4	99.39	3.17	58.45	1.86
1980	41.83	37.97	2327.8	88.39	2.82	−11.00	−0.35
1981	36.48	32.62	1971.0	64.29	2.01	−24.10	−0.75
1982	31.76	27.90	1569.5	43.79	1.41	−20.50	−0.66
1983	27.64	23.78	1573.1	37.41	1.16	−6.38	−0.20
1984	26.02	22.16	1727.5	38.29	1.11	0.88	0.03
1985	24.68	20.82	1565.9	32.60	0.92	−5.69	−0.16
1986	12.09	8.23	1930.9	15.89	0.44	−16.72	−0.46

Source: U.S. Department of Energy, Energy Information Administration, Annual Energy Review, 1986.

provides most of the detail needed roughly to distinguish rents from the payments needed to attract new resources into the oil and gas extraction.[5]

The *Survey* indicates that in 1974 the current and capital expenditures of the oil and gas extraction industry exclusive of leases were about $23.1 billion (in 1982 dollars). The equivalent figure for 1973, when production could not have been significantly affected by the oil price increases at the end of the year, was $18.1 billion, so there had been an increase of $5 billion. This is less than .2 percent of GDP in 1974.

---

[5] The cost in foregone output of the resources shifted into the domestic oil production is triangle *abc* in Figure 1, not the rectangle from *ab* up to the domestic producers' price. All resources that do not have an infinitely elastic supply to this industry normally earn payments beyond those necessary to keep them in this industry; factor payments to workers or firms with skills or products specialized to this industry are as likely to contain rents as those for land or natural resources. We may nonetheless keep rents from causing a large overestimation of the opportunity cost to society of resources shifting into domestic oil and gas extraction by leaving out the cost of leases of land or mineral rights and also the profits or residual incomes of the companies engaged in oil and gas extraction.

This sum is so small in relation to the loss from the productivity slowdown that no refinements on the supply side would seem necessary. But the true figure is even smaller. One reason is that some of these expenditures were capital expenditures that were allocated to this industry to obtain a flow over future years. According to the Bureau of Economic Analysis's *Fixed Reproducible Tangible Wealth*, the net investment in reproducible tangible capital stock in the natural gas and oil extraction industry increased by $2.13 billion from 1973 to 1974. Only the normal rate of return on capital is a cost in 1974. If we assume this return was 10 percent, then subtracting out the cost of the net investment and adding in a tenth of this amount leaves us with $3 billion, which is .1 percent of 1974 GDP. Similar calculations for the years after 1973 are given in Table 2.[6]

From Table 2 it is clear that all of the extra resources involved in oil and gas extraction in each year in the 1970s were, except in 1979, less than .3 percent of GDP, and that the year-to-year increase in these costs (see Column J) averaged only .06 percent of GDP over the 1970s. In the 1980s the value of the resources devoted to domestic oil and gas extraction was much greater. Still, all of the extra resources devoted to this industry even in the peak year of 1981 amounted to only 1.5 percent (and the increase over the prior year only .7 percent) of GDP. After 1981 these year-to-year figures are often negative and therefore favorable to U.S. economic performance.

Though these figures from Table 2 on the value resources transferred into domestic oil extraction are broadly confirmed by independent data,[7] they nonetheless probably tend to overstate the supply side losses in productivity. This is because no account has been taken of the extra oil these resources produced; the extra oil produced, valued at pre-OPEC prices, should be subtracted from the cost of the

---

[6] Note that the loss of output to the rest of the economy from the capital shifted to the oil and gas extraction industry would be the return on the extra capital invested in *all years since the oil shock* at the end of 1973, because all of this capital would in the absence of the oil shock presumably have been earning the normal rate of return in other parts of the economy. Table 2 is constructed on this basis (see Column F).

[7] The independent data are from the BEA Input-Output Tables for 1972, 1977, and 1981. We see from Column A of Table 2 that according to the *Survey* total real expenditures exclusive of leases increased by $15.7 billion between 1973 and 1977. If one takes only the BEA figures on expenditures on intermediate goods, wages, and indirect business taxes, (that is, excludes everything that the BEA calls "property-type income") to minimize the inclusion of rents, the increase in the value of resources in oil and gas extraction from 1972 to 1977 was $10.9 billion, or about a third less. But if one excludes only the residual income categories of corporate and unincorporated profits from the BEA figures, then the increase in resources was $19.2 billion, or nearly a fourth more than the Census figures. Column A of Table 2 reveals that between 1973 and 1981 total real expenditures measured by the *Survey* increased $73.9 billion. Between 1972 and 1981 the BEA figures show that this increase was slightly less ($69.2 billion) if all property type income is excluded and significantly more ($82.5 billion) if only profits are excluded. Between 1977 and 1981 *Survey* expenditures increased by $58.2 billion, whereas BEA expenditures increased by almost exactly the same amount ($58.3 billion) if all property-type income is excluded, but by $63.3 billion if only profits are excluded. For the BEA data see Interindustry Economics Division (1984), Table 1, page 52; and Ritz and others (1979) Table 1, p. 62.

Table 2
Resources Devoted to Oil and Gas Extraction
(in billions of 1982 dollars)

Year	*Total Expenditures Exclusive of Leases*	*Investment (Real Capital) Expenditure*	*Non-Capital Expenditure (A)–(B)*	*Increase in Non-Capital Expenditure from 1973 level (7.95)*	*Increase in Real Capital Expenditure from 1973 level (10.15)*
	(A)	(B)	(C)	(D)	(E)
1973	18.10	10.15	7.95	0.00	0.00
1974	23.07	12.29	10.78	2.84	2.13
1975	28.06	15.22	12.85	4.90	5.06
1976	30.35	15.87	14.48	6.54	5.72
1977	33.82	19.92	13.90	5.96	9.77
1978	37.45	24.15	13.30	5.36	13.99
1979	40.89	27.81	13.08	5.13	17.66
1980	59.91	33.07	26.84	18.89	22.92
1981	91.99	46.28	45.71	37.77	36.13
1982	82.11	45.15	36.96	29.01	35.00
1983	n/a	n/a	n/a	n/a	n/a
1984	70.37	27.87	42.51	34.56	17.71
1985	66.15	24.32	41.83	33.88	14.17

resources shifted into the oil extraction industry to get the net supply side loss of productivity.

When the costs on the demand side[8] of substituting away from more expensive oil are added to the costs on the supply side, we have essentially all of the losses in productivity due to higher prices of imported oil. We can determine approximately how large the cost illustrated by the area *gfh* in Figure 1 is by estimating the size of the drop in petroleum consumption per dollar of real GDP after the oil price shocks. Ultimately, this estimate could be made with a full-scale econometric analysis of the price and income elasticities for petroleum. But the only purpose now is to determine whether the substitution effects of the higher prices of imported oil can explain any large part of the productivity slowdown, so the approach here will be to make extreme and transparently simple assumptions that immediately provide a compelling upper bound on the size of these effects.

---

[8] The increase in prices of imported oil reduced real incomes in the United States and also made the prices of other forms of energy rise. The first development shifted the demand curve for oil to the left, the second to the right. The partial equilibrium analysis here accordingly offers only an approximation. The conclusions would not, however, be changed by taking the aforementioned shifts in the demand curve into account.

Table 2
Continued

	Sum of Increase in Real Capital Expenditure after 1973	*10% of Column (F)*	*Total Extra Resources Devoted to Industry after 1973*		*Increase in Column (H) over Prior Year as % of GDP*
			in bill. $	*as % of GDP*	
	(F)	(G)	(H)	(I)	(J)
1973	0.00	0.00	—	—	—
1974	2.13	0.21	3.05	0.11	0.11
1975	7.20	0.72	5.62	0.21	0.10
1976	12.91	1.29	7.83	0.28	0.08
1977	22.68	2.27	8.22	0.28	0.01
1978	36.67	3.67	9.02	0.29	0.03
1979	54.33	5.43	10.56	0.34	0.05
1980	77.25	7.72	26.62	0.85	0.51
1981	113.38	11.34	49.10	1.54	0.70
1982	148.37	14.84	43.85	1.41	-0.17
1983	n/a	n/a	n/a	n/a	n/a
1984	166.09	16.61	51.17	1.48	n/a
1985	180.26	18.03	51.91	1.46	0.02

n/a: not available.

GNP deflators are used to calculate above data in billions of 1982 dollars.

Sources:

1. U.S. Department of Commerce, Bureau of Economic Analysis, Fixed Reproducible Tangible Wealth in the U.S., 1987.
2. U.S. Department of Commerce, Bureau of Census, Current Industrial Report, Annual Survey of Oil and Gas, 1974, 1975, 1979, 1982.
3. American Petroleum Institute, Survey on Oil and Gas Expenditures: for Exploration, Developing, and Production, 1984, 1985.

Conveniently, U.S. petroleum consumption had stabilized by 1955 at just over two barrels of petroleum per thousand dollars of real GDP (in 1982 prices). This ratio went down a trifle between 1955 and 1970 and (perhaps because of pollution control regulations) then went up to 2.3 barrels per $1,000 of GDP in the first three years of the 1970s. So let's generously suppose that oil consumption would have been higher than ever—2.5 barrels per thousand dollars of GDP in 1982 dollars—throughout the oil crises, had that crisis never occurred. This assumption is embodied in Part A of Table 3. For safety, let us also make the alternative assumption that there was a continuing upward trend of .3 percent per year, a figure taken from the slope of a regression with time from 1955 through 1973 as the independent variable and the

Table 3A
Substitutions on the Demand Side

			Substitution Costs			*Annual Increase in Substitution Costs*	
			No-Gain Substitution				
Year	*Decrease in Consumption Per 1000$ GDP from assumed level (2.50 barrels)*	*Substitution Estimates (million barrels)*	*in billions of 1982 dollars*	*as Percentage of GDP*	*Assuming Linear Demand Curve and 1/3 Consumer Demand 1/3 column (D)*	*No-gain as Percentage of GDP*	*Assuming Linear Demand Curve and 1/3Consumer Demand 1/3 column (F)*
	(A)	(B)	(C)	(D)	(E)	(F)	(G)
1974	0.24	656.5	10.81	0.40	0.13	0.37	0.12
1975	0.27	707.4	10.05	0.38	0.13	- 0.03	-0.01
1976	0.21	593.9	8.65	0.31	0.10	-0.05	-0.02
1977	0.20	576.1	8.54	0.29	0.10	-0.00	-0.00
1978	0.26	802.2	11.25	0.37	0.12	0.09	0.03
1979	0.35	1085.3	36.99	1.18	0.39	0.82	0.27
1980	0.51	1585.3	60.20	1.92	0.64	0.74	0.25
1981	0.66	2122.1	69.22	2.17	0.72	0.28	0.09
1982	0.71	2202.5	61.45	1.97	0.66	-0.25	-0.08
1983	0.78	2519.1	59.91	1.85	0.62	-0.05	-0.02
1984	0.83	2857.8	63.34	1.84	0.61	0.10	0.03
1985	0.88	3129.3	65.16	1.84	0.61	0.05	0.02
1986	0.88	3218.4	26.48	0.73	0.24	- 1.06	-0.35

Sources: U.S. Department of Energy, Energy Information Administration, Annual Energy Review, 1986.

natural log of oil consumption per dollar of real GDP as the dependent variable. This assumption is embodied in Part B of Table 3.

In 1974, on the first of these assumptions, the zero-price-elasticity consumption of oil would have been 656.6 million barrels more than the actual consumption of 6077.2 million barrels. Even assuming that every barrel of reduction in consumption cost so much that nothing would have been lost by buying another barrel at the world price, the loss was $10.8 billion, or only .4 percent of 1974 GDP. On the second assumption, this loss was only just over .1 percent of GDP in 1974 and always less than .2 percent of GDP until the second oil shock. These costs were higher during and after 1979. This is not only because the second oil shock raised prices further; more substitutions occur as time goes on, and this second assumption assumes that if prices had not increased consumption would have increased in each successive year. Even so, the year-to-year increases in substitution that are most relevant are never much more than .7 percent of GDP (in 1980) and in all years except 1979–81 are either no more than .1 percent of GDP or else negative.

Given that at least a third of the petroleum was used by consumers for private vehicles and home heating[9] and accordingly did not affect measured productivity (though the higher price reduced the consumers' welfare), these numbers must be reduced by at least a third. In addition, the foregoing numbers unrealistically assume that all of the substitutions were "no gain" substitutions that brought no net benefits whatever—that they cost the whole rectangle *gekh* in Figure 1, or exactly as much as it would have cost to import all the oil they replaced. In fact, there were surely some nearly costless substitutions at the margin of indifference. Suppose that the demand curve in the relevant range was linear and (to err on the high side) that domestic buyers paid the higher world price. Then the relevant triangle will be half as large as the rectangle just referred to. This means that the foregoing numbers would need to be reduced by one-half again, which in combination with the foregoing allowance for consumer use of petroleum will make the estimate a third of what we calculated earlier and showed in the "No-gain Substitution" columns of Tables 3A and 3B.

Thus the productivity losses from the demand side were very small indeed. It was the productivity slowdown in the 1970s that generated most of the literature attributing the problem to oil, but before 1979 the losses on the demand side were never more than .1 percent of GDP.

When the direct productivity losses on both sides of the oil market are added together, they are seen to be very much too small to explain the productivity slowdown of the 1970s or even any large proportion of it. In the years before 1979 the annual productivity losses on the supply side are only .1 percent of GDP and the losses on the demand side are often even smaller than that. The many well-known estimates showing huge substitution losses from higher prices of oil or energy in the 1970s are wrong, and by a large multiple.

The claim that substitution away from oil was the main cause of the productivity slowdown is slightly less farfetched for the 1980s. The peak year for losses was 1981, chiefly because of the large shift of resources into domestic oil and gas extraction that year. But even then the direct productivity loss is significantly less than 1 percent of GDP.[10]

Some economists may think that advanced techniques will demonstrate that the direct impact of the substitutions induced by the higher prices of oil do explain the productivity slowdown. I hope the transparency of the foregoing calculations will convince them that this supposition is about as reasonable as supposing that more sensitive biological instruments will show that mice devour elephants.

---

[9] See *Annual Energy Review* 1986, Table 56, p. 123 and Darmstadter and others (1983), Table A-1, p. 50 The business sector made capital investments to provide fuel-efficient cars and other goods for consumers, but these investments counted as final output in the national accounts and did not generally affect measured output except to the extent they were incorrectly reported as current business expenses.

[10] If we consider the total loss in a given year from the whole of the price increase since 1971-72 the picture is the same. By 1981 the total direct productivity loss from all of the substitutions on the demand side was about .7 percent and the loss on the supply side was about 1.5 percent of GDP, for a total of more than 2 percent. But if Denison's residual productivity had continued to increase at the rate of 1.4 percent per year after 1973 productivity would have increased by nearly 12 percent by 1981, so the loss would, even in this worst year, explain less than a fifth of the productivity slowdown.

Since the cost of energy is such a small part of GDP, some economists like Arthur Okun, George Perry, Edward Denison, and William Nordhaus long ago argued that it could not explain the productivity slowdown. Other economists, aware that the two phenomena began in the same year, hypothesized that energy costs had a huge impact because they influenced some other variable large enough to account for the productivity slowdown. Energy and capital, it was said, were complements, so that higher energy prices slowed down investment, an obviously important source of economic growth. But the level of investment remained much too high to sustain belief in this hypothesis and in any case a reduction of an input such as investment need not lead to a productivity slowdown. Then there was the hypothesis that high energy prices made much of the old, fuel-inefficient capital stock obsolete, and that conventional measurements of the capital stock did not capture this obsolescence, so the apparent productivity slowdown could be largely an artifact of an unmeasured decline in the capital input. But this hypothesis also was undermined, in part by the finding that the relative prices of used energy-inefficient equipment had not changed much after the oil shocks (Hulten, Robertson, and Wykoff, 1987). Then econometrics apparently came to the rescue with estimates suggesting that energy prices influenced the rate of innovation and by that route caused the productivity slowdown. Since the oil shocks and the

Table 3B
Substitutions on the Demand Side

Year	*Petroleum Products Consumption (million barrels)*	*Hypothetical Consumption (million barrels)*	*Actual Consumption per 1000$ GDP (barrels)*	*Hypothetical Consumption per 1000$ GDP (barrels)*	*Substitution Estimates (million barrels)*
	(A)	(B)	(C)	(D)	(E)
1970	5365.5		2.24		
1971	5551.7		2.25		
1972	5991.4		2.32		
1973	6318.2	6318.2	2.33	2.33	0.0
1974	6077.2	6297.1	2.26	2.34	219.8
1975	5956.8	6253.5	2.23	2.35	296.7
1976	6390.4	6576.4	2.29	2.35	186.0
1977	6727.0	6900.1	2.30	2.36	173.2
1978	6880.3	7283.7	2.24	2.37	403.4
1979	6756.2	7460.0	2.15	2.38	703.8
1980	6244.0	7473.9	1.99	2.39	1230.0
1981	5861.9	7647.9	1.84	2.39	1786.0
1982	5584.5	7484.8	1.79	2.40	1900.3
1983	5559.0	7791.2	1.72	2.41	2232.3
1984	5757.2	8337.7	1.67	2.42	2580.6
1985	5741.5	8614.8	1.62	2.43	2873.3
1986	5891.1	8877.1	1.62	2.44	2986.0

productivity slowdown began in the same year, econometric estimates with annual data can easily suggest a causal connection between the two. But if the estimating equations are misspecified, the connection may be spurious. I shall return to this issue after introducing some concepts that make a better specification possible.

Where's the Boom?

History has been unkind to the argument that the oil shocks caused the productivity slowdown. The collapse of oil prices in 1986 rivaled the price rises of the 1970s. Those who believe that oil was the villain then need to explain what one of them, fair-mindedly, called "The Boom that Wasn't" (Mork, 1988). If oil caused the slowdown and the economy exhibited normal symmetry, productivity growth should have returned to pre-1973 rates. In extenuation, it is said (at least when short-term growth is at issue) that *decreases* in oil import prices are *also harmful* to the American economy because they induce costly reallocations of resources. In fact, if the costly reallocations are rational, they must make the net gains from cheaper oil imports even larger.

Though the timing of the oil shocks appears to be only slightly wrong for the "substitution against oil" hypothesis at the beginning and wildly wrong at the end, the timing is in fact worse at the outset for the familiar hypothesis that substitution against

Table 3B
Continued

Year	*Substitution Costs*			*Annual Increase in Substitution Costs*		
	No-gain Substitution		*Assuming Linear Demand Curve and 1/3 Consumer Demand*	*No-gain Substitution*		*Assuming Linear Demand Curve and 1/3 Consumer Demand*
	in billions of 1982 dollar	*as Percentage of GDP*	*1/3 of column (G)*	*in billions of 1982 dollar*	*as Percentage of GDP*	*1/3 of column (J)*
	(F)	(G)	(H)	(I)	(J)	(K)
1973	0.00	0.00	0.00	—	—	—
1974	3.62	0.13	0.04	3.62	0.13	0.04
1975	4.21	0.16	0.05	0.59	0.02	0.01
1976	2.71	0.10	0.03	−1.50	−0.05	−0.02
1977	2.57	0.09	0.03	−0.14	−0.00	−0.00
1978	5.66	0.18	0.06	3.09	0.10	0.03
1979	23.99	0.76	0.25	18.33	0.58	0.19
1980	46.70	1.49	0.50	22.72	0.73	0.24
1981	58.26	1.82	0.61	11.55	0.36	0.12
1982	53.02	1.70	0.57	−5.24	−0.17	−0.06
1983	53.09	1.64	0.55	0.07	0.00	0.00
1984	57.20	1.66	0.55	4.11	0.12	0.04
1985	59.83	1.69	0.56	2.63	0.07	0.02
1986	24.57	0.67	0.22	−35.26	−0.97	−0.32

Source: U.S. Department of Energy, Energy Information Administration, Annual Energy Review, 1986.

oil accounted for the productivity slowdown. Consider the earlier analysis here of the time-profile that the substitution costs should have and the patterns in Tables 2 and 3. It takes time to substitute new fuel-efficient technologies for old ones and the tables show that the losses in the first years after the first oil shock were minute, but that these losses were somewhat higher by the end of the 1970s. Some of the new capital chosen in response to the second oil shock was still being put in place in 1986, when oil prices were falling, and the oil-usage to real GDP ratio was then far lower than it had ever been in the 1970s. The substitution effects of the earlier oil price shocks do help in explaining why productivity has not recovered more in the 1980s; the main mistake was to attribute the productivity slowdown that started in 1973 to these substitutions, as many of the best-known studies did.

Another Approach: Collective Action and Institutional Rigidities

Though the vastly larger *total* costs of the oil shocks will, *in combination with other factors*, play a role in the explanation offered here of the productivity slowdown, we must first consider a phenomenon that is ignored in all but some of the most recent economic analysis: collective action. We need to consider organizations or collusions that lobby the government or combine in the marketplace to influence prices or wages —professional associations, labor unions, farm organizations, trade associations, oligopolistic collusions, and any other coalitions that attempt to shift the distribution of income in their direction.

The government favors and the monopolistic or monopsonistic prices or wages obtained by combination are collective goods, analytically speaking, in that the benefits automatically go to every firm or individual in some industry or category; a tariff or tax loophole favors every producer in some group and cartelization and collusion raises the price or wage for every seller in the relevant market. Thus collective action requires, at least for large groups, "selective incentives"—punishments and rewards that are applied to individuals according to whether they act in the group interest. Though the only widely known selective incentives are the union shop and the picket line, all large organizations for collective action get most of their membership through selective incentives. Small groups, such as the few large firms in a concentrated industry, may organize without selective incentives, since each member gets a significant fraction of the benefit of any action it takes in the interest of the group and strategically motivated cooperation may also occur. The logic that generates the foregoing assertions is set out in my *Logic of Collective Action* (1965).

Since all collective action must overcome a public good problem, even those groups that are small or that have the potential to work out selective incentives will be able to organize only when they have good leadership and favorable circumstances. Once financial security has been achieved, however, an organization will normally keep going indefinitely unless it is destroyed by violence of repression. It follows that long-stable societies will have accumulated many of these special interest groups whereas societies that have been unstable will have relatively few.

Suppose the clientele of an organization for collective action normally receives 1 percent of the GNP. The members of the organization will then bear, on average, 1 percent of any losses in the society's efficiency or capacity to innovate that their coalition brings about, but they will obtain virtually all of the benefits of their action. They then have an incentive to press for special privileges and monopoly until the GNP falls by a hundred times as much as they obtain. A society dense with these coalitions concerned with distributional struggle instead of production is like a china shop filled with wrestlers battling over its contents and breaking more than they carry away. Though the foregoing argument is incomplete, the logic has been spelled out and its implications tested against all of the most striking variations in national economic performance in modern times, in my book on *The Rise and Decline of Nations* (1982).

Some testable implications of the argument are immediately apparent. If collective action is inherently difficult and can be worked out only in fullness of time, and if narrow coalitions for collective action have incentives to engage in distributional struggle with virtually no concern for its cost to society, then long-stable societies should be less efficient and dynamic than comparable societies. that have had time to accumulate organizations for collective action.

As the theory predicts, they are. To take one example, the society that has had the longest period of stability and immunity from invasion and institutional destruction is the United Kingdom. This is also the country with the poorest overall postwar economic performance of all of the major developed democracies. The theory also predicts that if repression and defeat in war have destroyed the distributional coalitions in a society, then after a free and stable legal order is established the society should grow surprisingly rapidly. The "economic miracles" after World War II of Germany, of Italy, and of Japan and its former colonies are consistent with this implication.

Coalitions for distributional struggle are also undercut when many parochial governments with their own trade barriers are replaced by far larger jurisdictions with internal free trade; local monopolistic combinations such as guilds are undercut by outside firms and lobbies of a different order of magnitude are needed to influence the integrated jurisdiction. This helps explain the growth of previously-poor Germany after the Zollverein and national unification in the middle of the 19th century, the growth of once-backward Japan after the Meiji revolution of 1867–68, the growth of Holland during its Golden Age in the 17th century, the growth of Britain during the Industrial Revolution from about 1760 to about 1840, and the commercial revolutions in England and France in the 16th century.

The first of the five outstanding facts with which this article began was the slower growth of the United States relative to most comparable countries in the first quarter century after World War II. Given the foregoing intellectual framework, this is not puzzling. The United States did not begin the years after World War II with anything like the clean institutional slates of the defeated countries. The United States also did not experience the emasculation of its trade associations and oligopolistic collusions of manufacturing firms that France and some other protectionist countries experienced with "jurisdictional integration" via the Common Market at the end of the 1950s. The aging American economy accordingly suffers from an "institutional sclerosis."

Since World War II, the Continental European countries and Japan have enjoyed over forty years of stability and during that time they have gradually acquired more distributional coalitions. As some European economists have put it, the continent of Europe is coming to suffer from "Eurosclerosis." As the theory would predict, their lead in growth rates has diminished as well, though part of the reason is also their lesser opportunities for catch-up growth as they got closer to the U.S. level.

The same logic applies to old industries; in a long-stable society, the older an industry is, the more time firms and workers in the industry will have had to overcome the difficulties of collective action and the greater the obstacles to efficiency and innovation. Peter Murrell (1983) has shown that, as the theory predicts, new British industries after World War II performed nearly as well as their West German counterparts, whereas old British industries performed much worse than those of comparable age in Germany. Thus we have an explanation for the second outstanding fact—the tendency for the long-stable United States to perform badly in many old mass production industries in which it once led the world at the same time it apparently does relatively well in new and often high-technology industries. There is, moreover, conspicuous evidence that some old industries such as steel or autos have had considerable experience of combination or price leadership (as in the "Pittsburgh Plus" system) and that cartelization of the labor forces in these industries had led to wages that have often been two-thirds or more above the average of American manufacturing. By contrast, in new and especially high-technology industries cartelization and lobbying by firms has less often been a factor and the labor force is usually allocated by competitive markets.

The logic also applies within regions of a country, when account is taken of which activities are mobile and which are not. When an industry is constrained by natural resources to a particular region of the country, a lobby or collusion on behalf of the firms in the industry can increase income in that region at the expense of the rest of the country. But when labor or other factors supplied to footloose manufacturing industries cartelize and raise costs, these industries move to less organized areas, with the result that the relatively cartelized region loses much more than the nation does and relatively uncartelized regions gains. Thus the theory predicts that the regions with disproportionate cartelization of factors used by mobile manufacturing industries will lose population, capital, and total income to the areas with more competitive markets for their inputs (Olson, 1983). Though data are lacking on the degree of organization of some inputs for footloose industry, there are good data by state on labor unions; they have been less strong in the recently settled West and in the South, where black and racially integrated organizations were essentially prohibited until the civil rights legislation of the mid-1960s. The percentage of the labor force unionized by state has a strong and statistically significant negative correlation with growth of total income and a strong positive correlation with unemployment rates (Kendix, 1988; Olson, 1982, 1986; Summers, 1986). Some of the migration of labor is paradoxically from high-wage to low-wage areas (Hulten and Schwab, 1984) and this is further evidence of the disequilibria caused by cartelization.

Thus the theory that has just been evoked appears to explain the first three outstanding facts about postwar U.S. growth. Though further research is needed, the great majority of the statistical tests and other studies of the theory that have been done so far support it (see, for example, Chan, 1987; Choi, 1983; Vedder and Gallaway, 1986; Paloheimo, 1984; Weede, 1986, 1988; and the collections of assessments in *International Studies Quarterly* 1983; Mueller, ed., 1983; *Scandinavian Political Studies*, 1986).

Any explanation of the fourth outstanding fact must take account of the relatively sudden worsening of the slowdown in growth and productivity. Obviously, no gradual process, such as the accumulation of distributional coalitions that has just been evoked could, by itself, explain a sudden change. Sharper changes, such as macroeconomic fluctuations, could. The productivity slowdown has, however, lasted nearly a decade and a half, and this is too long to be explained by the conventional macroeconomic theories. I shall therefore outline an approach to macroeconomics that grows out of the theory evoked above, and then show how this approach, in combination with macroeconomic insights of other economists, can help explain both the productivity slowdown and the dramatic deterioration in macroeconomic performance that occurred at the same time.

The Macroeconomic Implications

The approach to macroeconomics suggested here begins with the incentive for mutually advantageous transactions between buyers and sellers. Were it not often countervailed, this incentive would make all markets clear and thus eliminate depression and involuntary unemployment. All schools of thought agree that in purely competitive markets there will be transactions until the supply and demand curves intersect, that the marginal costs of sellers are then equal to the value buyers put on a marginal unit of the good, and that any additional transactions would be socially inefficient. But even if there were only one buyer and one seller, the two bilateral monopolists would have an incentive to bargain until they had made all mutually advantageous trades, and when this has happened the seller's marginal cost must again be equal to the value of a marginal unit to the buyer. When there is either a single firm with monopoly power or monopolistic competition, there cannot be Pareto-efficiency, but when the seller's marginal revenue just equals its marginal cost the firm wants no larger output, and at the price that is charged the consumer wishes to buy no more, so again markets clear and there is none of the disequilibria or involuntary unemployment associated with depressions.

The gains from trade and the incentives to transact until markets clear are present irrespective of the level of aggregate demand (though nominal prices and wages obviously depend on the level of nominal spending) and they are present in labor markets as much as product markets. No worker could reasonably be described as involuntarily unemployed unless he valued his time at less than it would be worth to some employer, but whenever this is true both the unemployed worker and the employer could gain from making a deal, so involuntary unemployment would seem to be impossible.

But if something happens, it is possible. For most of us, the macroeconomically ideal outcomes that have just been described run against all experience. The reason, I have argued (Olson, 1984), is that in our society many of the mutually advantageous transactions that would eliminate unemployment and idle capacity are prevented by parties other than the buyer and the seller who would gain from the transactions.

Who has an incentive to block such transactions? Whenever the firms or workers in an industry have acquired the capacity to act collectively, they can profit by establishing a noncompetitive price. Whenever any such price has been established, there must be potential transactions at a lower price that would profit both individual sellers and individual buyers. If these transactions occur, they eliminate the supracompetitive price. Thus no coalition can ever obtain a noncompetitive price through collective action without blocking transactions itself or using its lobbying power to get the government to block them. Whenever a coalition does this it reduces output and employment. Thus the theory predicts that the accumulation of organizations for collective action in stable societies leads to a secular increase in the "natural" or "non-accelerating inflation" rate of unemployment.

The decision about precisely what supracompetitive price a coalition should set has to be decided collectively. When numbers are small, as in an oligopolistic collusion, the decisions are normally made by bargaining until there is consensus, which is necessary because dissenting firms would undercut the price. Yet the conflict of interest over who gets to sell how much at the profitable supracompetitive price makes consensus difficult to achieve. When the membership is large, decisions have to be made according to constitutional rules. But a significant minority can still undermine the market power or lobbying clout of an organization, so it must have procedures and patterns of consultation that insure that there is general support for major decisions. Whether important decisions are made by bargaining-to-consensus or under by-laws, they are normally made slowly. Thus, the prices and wages set by distributional coalitions are changed only periodically. In collective bargaining in the United States, to take an extreme example, decisions about wage levels are normally made only once every three years.

Since there is a divisive conflict of interest about who gets to sell how much at the supracompetitive price, most coalitions set prices rather than quantities and let impartial mechanisms, such as seniority rules or customer preferences, apportion quantities. The preference for price-fixing over quantity-fixing, in combination with slow decision-making, means that the prices that are too high to clear markets are also sticky (Olson, 1982).

The typical member of the large special-interest organization, like the typical citizen, is normally "rationally ignorant" (Downs, 1957; Olson, 1982). The probability that an average voter will change the course of a large organization or of a country is vanishingly small, and if beneficial changes do occur they are shared by the entire country or group. Thus the expected value of information about the affairs of governments or large organizations to an average individual is minuscule and he will not sacrifice much to get it.

One result is that the choices of distributional coalitions and of the governments of democracies will be overwhelmingly influenced by the information that is automatically or costlessly acquired through individual experience. Even some pertinent individual experience will be forgotten, so experiences that are easily remembered because they were traumatic, repetitive, or recent play a disproportionate role. If Germans, for example, have been unforgettably affected by a hyperinflation, that will affect later choices in Germany more than it affects decisions in other societies, even though people in other countries could study German experience. In the same way, the great 1929–41 depression in the United States had a traumatic impact on the decisions of those who lived through it; in their predictions, they gave it even more salience than it deserved. The typical person voting about collective goods will similarly often assume that the future will be like the recent past because that is often what he remembers.

The theory that has just been impressionistically described predicts that in a sclerotic economy purely nominal surprises have major real consequences and that the impact of real shocks will be multiplied. Consider first *unexpected* inflations, disinflations, or deflations. If the memorable experience of the memberships of coalitions does not include high inflation, these groups will set nominal rather than indexed prices. Since more distributional coalitions exercise monopoly than monopsony power, unexpected inflations will temporarily make the prices obtained by lobbying or by cartelization lower, in relation to the freely fluctuating prices in unorganized sectors, than they would normally be. Thus the social losses from monopoly and the extent to which prices and wages are above market clearing levels will be temporarily diminished and the productivity of the economy will improve for a time.

Unexpected deflations and disinflations have the opposite effects. The prices in the sectors controlled by coalitions are temporarily even farther above competitive and market-clearing levels than these coalitions want them to be. This increases the extent of unemployment and idle capacity (Olson, 1984).

Admittedly, there are always some sectors of a society that are not organized and unemployed resources can move to these sectors, but the migration further lowers the returns in these unorganized sectors. Identical resources then earn vastly different rates of return depending on whether they are organized or crowded into open sectors. Though some employment, such as selling apples on street corners, may be available, the returns when a sclerotic economy confronts unexpected deflation and disinflation are vastly below those in organized sectors. This increases the degree of monopoly, the extent of policing of transactions and innovations that endanger supracompetitive prices, the amount of socially unproductive searching and queuing, the extent of socially-inefficient regulation, and other forms of rent seeking (Olson, 1982).

Now suppose a surprise is real rather than nominal. Whenever any workers, members of a profession, or suppliers of any factor of production are organized, the optimal supracompetitive price or wage for them to demand will depend, among other things, on their expectations of the growth of productivity and real income in *the economy as a whole. Normally, the greater the expected increase in real GNP, the higher the real price it will be optimal for organized factor-suppliers to demand. It follows that, if the real income of the economy increases more than the memorable* experience of a coalition of factor-suppliers leads them to expect, the price that they demand will be lower than it would have been had they had perfect foresight or instantaneous decision-making; it will not be as far above competitive levels as intended. It follows that *unexpected* economy-wide increases in productivity growth and real income increase *efficiency* and *reduce unemployment* in a sclerotic economy, and the favorable consequences of the surprise are accordingly *multiplied*. By the same logic, unexpected decreases in the rate of growth of productivity and GNP are magnified because they reduce efficiency and increase idle capacity in such an economy.

When a sclerotic economy enjoys any favorable shock, the proportion of new businesses and other risky ventures that succeed increases. These new ventures will not only directly magnify the increase in real income, but (since new firms, industries, and activities will not have been around long enough for the difficulties of collective action to be overcome) also reduce the society's degree of sclerosis. Similarly, an unexpected adverse development will be more severe and longer lasting because it raises the failure rate of productivity-increasing innovations and new ventures, and thereby also brings about a more rapid increase than would otherwise occur in the coalition density of the society. The magnification of unexpected shocks that occurs because shocks alter the degree of sclerosis affects both factor markets and product markets.

Thus unexpected real shocks, whether favorable or unfavorable, have larger and longer impacts on densely organized societies than on coalition-free societies. The multiplication of real shocks explained here suggests there is a need for real models of the business cycle. But the argument here does not support any real (or monetary) theory that assumes there is continuous market-clearing, or any model that concludes that unexpected monetary (or Keynesian) shocks cannot have real effects.

The Macroeconomic Divide and the Productivity Slowdown

The postwar United States had a lot of distributional coalitions and by the late 1960s or 1970s this was coming to be true even of the countries of continental Europe. Because high rates of inflation had not in general been part of memorable experience, most coalitions set prices in nominal terms. In the first two decades after World War II, there had been, because of unforgettable experience, widespread fears of another great depression. These fears made distributional coalitions cautious about setting prices that could not be borne if there were a depression. In fact, there was not any replay of the great depression but inflation instead. This wide divergence of expectations and reality made the degree of monopoly, the amount of socially unproductive queuing and searching, the extent of policing of transactions and innovations, the degree of rent-seeking regulation, and the extent of unemployment much less than they would otherwise have been. It therefore helped to make multifactor productivity increase at faster than normal rates.

The 1950s and 1960s were also a time of many favorable real developments that had no counterpart in memorable experience. The United States in the 1950s and 1960s also was in a position to make many investments and innovations that had not been exploited during the great depression or during World War II, the shortages that followed it, or sometimes even during the Korean war. The exploitation of these pent-up possibilities made productivity greater than it would normally be. There was an unusually rapid gain from reallocation of resources from agriculture to urban industries.

At the same time, the growth rates of the defeated economies were "miraculously" and thus also unexpectedly high. The phenomenally fast growth of France and some other continental European countries after the jurisdictional integration of the six nations that founded the Common Market was also unexpected. The United States also gained from the extraordinary growth of its trading partners; there was a rapidly increasing demand for U.S. exports, and especially for many manufactured products in which the United States then had a near-monopoly, and also rapid decreases in the real cost of most of those goods the country imported. These gains were magnified by a great liberalization of foreign trade and investment at the same time. Economies of scale helped favorable productivity developments feed on themselves. All of these and other factors made productivity growth substantially exceed the historical norm.

By all accounts these favorable developments were not expected, which implies that coalitions of factor suppliers did not receive prices as much above market-clearing levels as they would normally have obtained. This in turn generated the productivity-multiplying and unemployment-reducing results that the argument here predicts. The nice surprises with their favorable multipliers increased the success rates of new and thus coalition-free ventures and thereby strengthened the virtuous circle.

By the end of the 1960s and the beginning of the 1970s a completely different set of memorable experiences, and even a new post-depression generation, had emerged. The experiences of both the price level, on the one hand, and of the rates of growth of productivity and real incomes, on the other, were dramatically different than those of the depression years. The fear of another great depression, or even of major recessions, faded as the 1960s went on. As the memorable experience came to be of substantial and increasing rates of inflation, organized interests naturally demanded higher nominal prices in anticipation of further inflation. The historically unprecedented and continuing increases in productivity and real incomes also came to be part of memorable experience. By the late 1960s, each generation expected to be much better off than its parents had been and each few years were expected to bring the customary increases in real rates of return.

The exceptional increases in U.S. productivity, and the even more dramatic increases in some other countries, could not continue forever. Within a couple of decades after World War II, the previously neglected innovations had largely been exploited. The gain in real output from reallocating resources from farm to city had largely disappeared. The special gains from trade that the United States had enjoyed because it had a near-monopoly of the capacity to produce certain products and technologies disappeared as the rest of the world adopted its technologies. To be sure, *other countries correspondingly gained from their increasing capacity to compete with the United States and this country also gained from the growth of its trading partners.* But at about the same time the especially rapid growth of other countries, arising from exceptional opportunities for "catching up" and often also from clean institutional slates, was ending. The gains from increasing economies of scale then also diminished.

The developments that have been described did not by themselves make productivity growth stop or even make it fall below historical levels, but they did reduce rates of growth of real incomes. And there was nothing in the *experience* of the new generation to lead it to expect a return to historically normal rates of growth.

The ever-increasing rates of inflation also could not be sustained. Though the rising rates of inflation eventually came to be expected, they were never welcomed. There were political pressures, especially in the early 1970s, to bring the rate of inflation under control. Ominously, disinflationary policies and the shock of a return to normal levels of productivity were coming at essentially the same time.

Then, just on top of this inauspicious conjuncture, came the oil shock, which was also unexpected. Note that it is the *much larger total social cost of the oil shocks*, *not the tiny direct productivity-reducing substitution cost*, *that is relevant here.* As Table 1 reveals, the oil import bill increased by 1.3 percent of GDP in 1974. In 1979 the extra cost, over pre-1973 levels, of the oil the United States imported took 3.17 percent of the GDP in that year, and it took 1.86 percent of GDP to pay the increase in the bill over 1978. When we add on the costs of substitution in 1979 the total loss as compared with 1971–72 was somewhat less than 4 percent of GDP, and the increase over 1978 was more than 2 percent of GDP.

Note that all these things came together at a time when an organization for collective action setting prices or wages for the next couple of years could, on the basis of the memorable experience of its members, reasonably expect an increase in real earnings and also a continued increase in the rate of inflation. Thus experience suggested that price or wage increases should be set that would capture the customary increase in real rates of earnings, and also anticipate rising inflation, either through indexing to the cost of living or through nominal demands that included provision for the continually increasing rates of inflation. Yet at this time productivity was not only in the process of falling from its unprecedented levels, but attainable real incomes were also being significantly reduced by the higher prices of imported oil. The indexing to the cost of living that was a rational adaptation to nominal changes created inadvertent overpricing when the real price of imported oil went up. At the same time (in spite of some accommodation of monetary and fiscal policies to the higher oil prices) there was also a disinflationary restriction in the rate of growth of aggregate demand.

Given the coalition density, many prices and wages would have been too high for efficiency and market clearing even with perfect foresight. But decisions based on memorable experience led for a time to coalition prices and wages much higher than they would have been if the natural fall from exceptional levels of productivity growth and the oil shock had been foreseen and the harmful effects of each of them were therefore multiplied. The disinflationary aggregate demand shock came at essentially the same time, and each of the three shocks made the coalitions err in the same direction—the direction that multiplied their adverse effects on economic performance. Cartelistic, collusive, and lobby-influenced prices were for three *cumulative* reasons even higher than the coalitions wanted them to be. Though there is, unfortunately, no adequate measure of the extent to which oligopolies, professional associations, and lobby-inspired regulation raised prices above competitive levels, there are data on union–nonunion wage differentials and, as the theory predicts, they increased (see, for example, Summers, 1986). Presumably the degree of monopoly, the extent of prices above market-clearing levels, the amount of queuing, searching, and policing of transactions and innovations, and the amount of regulation and rent-seeking generally, increased. There certainly was, as the argument here implies, also a *simultaneous* deterioration in macroeconomic performance and in productivity growth.

But why hasn't the productivity slowdown and the mediocre economic performance completely ended by now? The previously discussed time profile of the substitution against oil is surely only part of the story. The same may be said for the tendency for adverse developments to reduce the number of successful new and thus coalition-free ventures. A more intriguing part of the story is probably evident from some independent but resonant models of "hysteresis" or "persistence" developed by Blanchard and Summers (1986) and by Lindbeck and Snower (1987). These authors point out that bad times reduce the number of "insiders" and that this normally makes it advantageous for insiders to choose higher wages than before, because the smaller the number of insiders the less the likelihood that a given unexpected shock will lead to unemployment for them. This in turn can make bad times persist. The idea has so far been applied only to the labor market, but with modification it could be applied in the spirit of this article to diverse markets in which there is collective action.

Though the argument in this article is consistent with all of the prominent evidence, we must remain cautious until the matter has been studied in great detail. The evidence must be explored at a much less aggregate level. Ideally, the power of the argument here should be tested with highly disaggregated econometric models such as some of those that have been impressively used to study the productivity slowdown. It also needs to be tested against detailed qualitative and historical studies of pertinent organizations and institutions.

The evidence that has already been presented is nonetheless sufficient to show that broader approaches to the study of the productivity slowdown are needed. Any econometric estimates of the causes of the productivity slowdown are seriously misspecified if the macroeconomic and institutional variables considered here are omitted. Independent checks of the leading econometric estimates (Norsworthy, 1983, 1987) and the calculations in this article also suggest that less austere specifications of econometric models are needed. The evidence that has been presented here is also sufficient to show that a narrow approach to macroeconomics is also insufficient. Economists cannot understand macroeconomic performance in aging economies without taking into account the incentives to obtain supracompetitive prices or wages through collective action, the resulting blocked transactions and nonclearing markets, *the slow decision-making inherent in collective decisions, the memorable experiences of the rationally ignorant, and the responses of organizations for collective action to both nominal and real surprises. This article has explained all five of the most prominent facts about the postwar U.S. economy with an extremely simple and parsimonious theory that has elsewhere explained other facts as well. It is most unlikely that so brief and simple a theory could account for so much if it did not focus on causally crucial phenomena.*

■ *The author is grateful for criticism and other help from Clopper Almon, Martin N. Baily, Ernst Berndt, Douglas Bohi, Christopher Clague, Joel Darmstadter, Edward Denison, Stanley Fischer, Charles R. Hulten, Dale Jorgenson, John Kendrick, Margaret B. McCarthy, Carl Shapiro, Joseph Stiglitz, Robert Schwab, Timothy Taylor, and Michael Toman. He is solely responsible for the faults of this article.*

References

Blanchard, Olivier, and Lawrence Summers, "Hysteresis and the European Unemployment Problem." In Fischer, Stanley, ed., *NBER Macroeconomics Annual, 1986*. Cambridge, Mass.: Massachusetts Institute of Technology Press, 1986.

Chan, Steve, "Growth with Equity: A Test of Olson's Theory for the Asian Pacific-Rim Countries," *Journal of Peace Research*, 1987, *24*, No. 2, 135–149.

Choi, Kwang, *Theories of Economic Growth*. Ames, IA: Iowa State University Press, 1983.

Darmstadter, Joel, et al., *Energy Today and Tomorrow*. Englewood Cliffs, N.J.: Prentice-Hall, 1983, Table A-1.

Denison, Edward F., *Accounting for Slower Economic Growth*. Washington, D.C.: The Brookings Institution, 1979.

Denison, Edward F., "International Transactions in Measures of the Nation's Production," *Survey of Current Business*, May 1981, *61*, No. 5, 17–28.

Denison, Edward F., *The Sources of Economic Growth in the United States and the Alternatives Before Us*. New York: Committee for Economic Development, Supplementary Paper No. 13, 1962, p. 245.

Downs, Anthony, *An Economic Theory of Democracy*. New York: Harper and Row, 1957.

Griliches, Zvi, "Productivity, R&D, and Basic Research at the Firm Level in the 1970s," *American Economic Review*, March 1986, *76*, 141–54.

Hulten, Charles R., and Robert M. Schwab, "Regional Productivity Growth in U.S. Manufacturing: 1951–1978," *American Economic Review*, March 1984, *74*, No. 1, 152–162.

Hulten, Charles R., James W. Robertson, and Frank C. Wykoff, "Energy, Obsolescence, and the Productivity Slowdown," June 1987, University of Maryland, unpublished manuscript.

Interindustry Economics Division, "The Input-Output Structure of the U.S. economy, 1977," *Survey of Current Business*, May 1984.

International Studies Quarterly, *27*, 1983, articles by Charles Kindleberger, Jan DeVries and Brian Barry.

Kendix, Michael, "An Investigation into the Effects of Labor Unions on Unemployment with Decentralized Union-Firm Bargaining Structures: Empirical Evidence for Britain, Canada and the United States," June 1988, University of Maryland, unpublished manuscript.

Klein, Burton H., *Germany's Economic Preparation for War*. Cambridge, MA: Harvard University Press, 1959, pp. 207–217.

Knorr, Klaus, *The Power of Nations: The Political Economy of International Relations*. New York: Basic Books, 1975. p. 19.

Lindbeck, Assar, and Dennis J. Snower, "Union Activity, Unemployment Persistence and

Wage-Employment Ratchets," *European Economic Review*, February/March, 1987, *31*, Nos. 1/2, 157–67.

Lindbeck, Assar, "The Recent Slowdown in Productivity Growth," *The Economic Journal*, March 1983, *93*, 13–34.

Mork, Knut Anton, "The Boom That Wasn't: Evidence From 1986–87 on the Macroeconomic Effects of Oil Price Fluctuations." Vanderbilt University, unpublished manuscript, 1988.

Mueller, Dennis C., ed., *The Political Economy of Growth*. New Haven: Yale University Press, 1983.

Murrell, Peter, "The Comparative Structure of the Growth of West German and British Manufacturing Industry." In Mueller, Dennis, ed., *The Political Economy of Growth*. New Haven: Yale University Press, 1983, pp. 109–131.

National Science Foundation, *National Patterns of Science and Technology Resources*. Washington, D.C.: 1986, #86-309, Table 5, 37.

Norsworthy, J. R., "Energy Prices, Technical Change, and Productivity Growth." In Schurr, Sam H., Sidney Sonenblum, and David O. Good, eds., *Energy, Productivity, and Economic Growth*. Cambridge, MA: Oelgeschlager, Gunn and Hain, 1983, pp. 155–77.

Norsworthy, J. R., "Comment." On "Jorgenson, Dale W., Frank M. Gollop, and Barbara M. Fraumeni, Productivity and Growth of Sectoral Output in the United States, 1948–79," *AEI Occasional Papers*. Washington, DC: American Enterprise Institute for Public Policy Research, January, 1987.

Olson, Mancur, "The Economics of Target Selection for the Combined Bomber Offensive," *Journal of the Royal United Service Institution*, November 1962, *CVII*, 308–14.

Olson, Mancur, *The Economics of the Wartime Shortage: A History of British Food Shortages in the Napoleonic War and World Wars I and II*. Durham, NC: Duke University Press, 1963.

Olson, Mancur, "American Materials Policy and the Physiocratic Fallacy," *Orbis*, Winter 1963, *VI*, 529–52.

Olson, Mancur, *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press, Harvard Economic Studies, *124*, 1965.

Olson, Mancur, *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven, CT: Yale University Press, 1982.

Olson, Mancur, "The South Will Fall Again: The South as Leader and Laggard in Economic Growth," Presidential Address to the Southern Economic Association, published in *Southern Economic Journal*, April 1983, *49*, No. 4, pp. 917–932.

Olson, Mancur, "Beyond Keynesianism and Monetarism," *Economic Inquiry*, July 1984, *XXII*, 297–321.

Olson, Mancur, "Maintaining a Healthy Business Climate: A Broader Perspective on the Rates of Economic Growth and of Unemployment in the Southern and Southwestern States," *Energy and the Southwest Economy: Proceedings of the 1985 Conference on Energy and the Southwest Economy*, Federal Reserve of Dallas, 1986, 271-304.

Paloheimo, Heikki, "Pluralism, Corporatism, and the Distributive Conflict in Developed Capitalist Societies," *Scandinavian Political Studies*, 1984, *7*, 17–38.

Ritz, Philip M., Eugene P. Roberts, and Paula C. Young, "Dollar-Value Tables for the 1972 Input-Output Study," *Survey of Current Business*, April 1979, *59*, 51–72.

Scandinavian Political Studies, March 1986, 9, various articles.

Solow, Robert, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, *XXXIX*, No. 3, 312–320.

Solow, Robert, *New York Times Book Review*, July 12, 1987, p. 36.

Summers, Lawrence H., "Why Is the Unemployment Rate So Very High Near Full Employment?" *Brookings Papers in Economic Activity*, 1986, No. 2, 339–86.

U.S. Department of Labor, Bureau of Labor Statistics, *Monthly Labor Review*. July 1974, Table 35, p. 119.

Vedder, Richard, and Lowell Gallaway, "Rentseeking, Distributional Coalitions, Taxes, Relative Prices and Economic Growth," *Public Choice*, 1986, *51*, No. 1, 93–100.

Weede, Erich, "Catch-up, Distributional Coalitions and Government as Determinants of Growth or Decline in Industrial Democracies," *British Journal of Sociology*, 1986, *37*, 194–220.

Weede, Erich, "Redistribution and Income Inequality in Industrial Democracies," Cologne University, Cologne, West Germany, unpublished manuscript, 1988.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 71-97

Tax Policy and Economic Growth: Lessons From the 1980s

Michael J. Boskin

The past seven years of U.S. economic policy and performance constitute a truly remarkable, if confusing, experiment. Few would have guessed in 1980 that inflation would fall rapidly to 4 percent and hold pretty steady during a lengthy recovery, that productivity growth would partially rebound, the national debt more than double, capital imports reach $150 billion in a single year, or that employment would continue to expand at rates envied by other advanced economies. Perhaps fewer still would have predicted two major and two minor tax reforms would be passed in six years, a flip-flop in depreciation schedules, and a top personal tax rate below 30 percent. While it is always easier to draw inferences concerning economic growth, its determinants, and policies affecting the determinants with long historical hindsight, this paper will attempt to provide information, evaluation, and conjecture to draw some tentative lessons about the effects of tax policies in the 1980s on long-run economic growth.

The good news from the 1980s is the remarkable achievement of substantially lower marginal tax rates, making the government a much smaller silent partner in many economic decisions. For example, the tax advantage of debt finance is reduced substantially with the much lower marginal tax rates. The bad news is that the nation could not afford all of the good news. The very large budget deficits add to the instability in the tax code, in addition to whatever other problems they cause. Moreover, in a country with the lowest saving and investment rates in the advanced world, the consequence of which may be to decrease our long-run growth potential and short-run trade competitiveness, the remaining anti-saving and investment features of our tax law, while probably less important in the long run than the U.S.

■ *Michael J. Boskin is Wohlford Professor of Economics and Director, Center for Economic Policy*

budget deficit, will eventually require serious attention. *The lesson of the 1980s is not that we should dramatically change the new tax laws but rather that we should build on their accomplishments: broaden the tax base to include more consumption while preserving low marginal tax rates and restoring carefully targeted incentives for saving and investment.*

Determinants of Growth

The power of compounding even modest increases in the growth rate is enormous. The United Kingdom, growing at only one percentage point per year less than the United States, France and Germany, transformed itself from the wealthiest society on earth to a relatively poor member of the Common Market in less than three generations.

As an illustration, consider two equally wealthy economies. In one, per capita income grows at 1.5 percent; in the other, at 2 percent. The more rapidly growing economy will be almost 30 percent more wealthy than the less rapidly growing economy in less than two generations, a difference in living standards that carries labels like "successful" and "sick," respectively. Thus, differences of *fractions of a percentage point* in the long run growth rate must be explained in attempting to assess growth performance. Increasing the growth rate (at minimal opportunity costs) by .2 or .3 percentage points is an enormous economic and social achievement.

Studies of economic growth usually attempt to decompose the rate of growth of real GNP (or some related output measure) into the contributions of various factors thought to explain it. These include such factors as increased labor input, increased capital input, improved resource allocation, and technical change.

The Empirical Evidence on Growth

As Table 1 demonstrates, real gross product and real product per hour worked in the United States have grown more rapidly in the period from 1981 through the first quarter of 1988 than from 1973 to 1981, but much slower than in the 1948–73 period.[1] Perhaps the quarter century after World War II should be thought of as an aberration, a special period of growth more rapid than is likely in the long run. What caused this history of rapid growth, a major slowdown, and the recent modest growth turnaround?

The usual growth accounting framework—based on the production function $F(K, L, t)$ where K is capital inputs, L is labor inputs, and t is time—asks how much could have been produced with the growth of inputs if technology was constant between two periods in time. The difference, the residual unexplained by input

---

[1] While the introduction of new products is a major source of increased economic well-being and is probably undervalued in the traditional GNP estimates, and numerous other measurement issues abound as well, I take the figures in Table 1 to be representative of the true trends. However, I have much sympathy with a Schumpeterian approach to growth emphasizing entrepreneurship. In the aggregate data, these gains are buried in investment, R&D, and other data.

Table 1
Output and Productivity Growth in the U.S. Economy
(average annual percentage rates of growth)

Year	*Real Gross Product*	*Real Product Per Employee Hour*
1948:IV–1973:IV	3.7	2.8
1973:IV–1981:III	2.2	0.7
1981:IV–1988:I	3.3	1.3

Source: U.S. Dept. of Labor

growth, represents the shift in the production function, which is taken to represent technical change, organizational efficiency gains, entrepreneurship, improved resource allocation, and so on. Many strong assumptions are usually made to draw such inferences: for example, constant returns to scale, profit maximization and competitive markets to allow use of factor shares as weights, no aggregation errors, capital malleability, no unmeasured inputs, no deviations of marginal products from normalized input prices, no scale or learning effects, no measurement errors in capital or labor input prices or utilization rates. Each of these issues is the subject of much research and debate. The sources and the quantitative determinants of economic growth remain a source of major controversy. However, I think the existing literature does support this mild conclusion:

> Lesson 1: The best empirical growth studies suggest that the rate of capital formation *and* technical change have been important determinants of long-run U.S. growth. These factors also help to explain differences in international productivity growth. Therefore, the effect of tax policy on investment *and* technical change is important in understanding the effects of tax policy on long-run growth.

When M. Abramovitz (1956) and R. Solow (1957) asked how much of the growth of output can be explained by the growth of inputs, their answer was precious little. Subsequent studies by E. Denison (1974), and others appeared to confirm this small contribution of increased labor and capital input to increased output. In particular, Denison (1979) places very little weight on the decline in the rate of growth of the capital-labor ratio as a contributing factor in the productivity growth slowdown. These authors placed a heavy emphasis on technological change as the *prime* determinant of economic growth. Jorgenson (1986) and others have questioned the Abramovitz-Solow-Denison finding, attributing a much greater share of postwar U.S. growth to increased capital input and much less to the unexplained residual "technical change."[2] In the work of Jorgenson and others, greater disaggregation and improved measures of labor quality seem to be a major reason for assigning greater weight to capital formation. As Lesson 1 states, I conclude from this evidence that both investment and innovation play a major role in determining productivity growth. In fact, as discussed below, I suspect investment and innovation are interdependent and that interdependence increases the elasticity of output with respect to capital inputs.[3]

Of course, saying that "innovation causes productivity" begs the question of what causes innovation. John Kendrick (1986) stresses the role of research and development. He estimates a typical lag of about six years between spending on R&D and its commercial application. He therefore attributes part of the decline in productivity growth in the 1973–81 period to a prior slowdown in R&D spending. Although R&D spending has increased recently, the lag involved in commercialization contributed to productivity growth being less strong in the 1981–86 period than otherwise.

Major efforts to incorporate R&D into the growth accounting framework (Griliches (1980) and his paper in this issue) conclude that changes in R&D spending explain only a modest part of the postwar productivity growth episodes. But R&D spending is not a sufficient proxy for innovation. Disembodied technical change and capital input growth (capital deepening and perhaps other effects discussed below) remain major foci for attention.

International growth rate comparisons also come to mixed conclusions on the relative (separate) importance of investment and technical change. Martin Baily and Alok Chakrabarti (1985) noted that while the absolute level of productivity in the United States is quite high, the rest of the world is catching up in productivity and industrialization. In the 1973–84 period, output per labor hour in the manufacturing sector grew 2 percent in the United States, but $3\frac{1}{2}$ times more rapidly in Japan and about twice as rapidly in France and Germany. R&D spending in these countries has risen faster than in the United States. Baily notes that the fraction of U.S. patents granted to foreigners had risen from 16 percent in 1960 to 41 percent by 1982. He stresses the importance of innovation and technical change to long-run growth, and with Denison, adopts the position that only 20 percent of U.S. productivity growth in this period was due to an increase in the capital/labor ratio.[4] The comparison with Japan, France, and Germany, however, could equally well highlight the very high cross-country correlation between investment and growth rates. These countries had investment and productivity growth rates two to three times those of the United States.[5] Kormendi and Meguire (1985) document the substantial contribution of differentials in investment rates to explaining cross-country differences in growth rates.

---

[2] Before proceeding to examine growth theory, it is worth noting that since the trough of the 1981–82 recession, cumulative real GNP growth in the United States has substantially exceeded that in France, Italy, Germany, and the United Kingdom. Only Canada (whose economy is tied closely to the United States) and Japan have fared as well as the United States in the 1980s.

[3] See P. Romer (1987) for a provocative argument stressing the importance of externalities generated by capital input. B. Arthur (1987) develops some interesting implications of positive feedback mechanisms in economics.

[4] Of course, the rate of growth of the capital/labor ratio slowed and this was one of the reasons why productivity slowed, but this reason was not the only one, nor by some estimates the most important one.

Distinguishing Permanent and Temporary Increases in the Growth Rate

Policies designed to alter the rate of economic growth tend to focus on enhancing technological advances, the quality of the labor force, and the level or rate of growth of capital per worker. It is important to note that the only way to raise the long-run growth rate permanently is to increase the rate of technical change or the rate of improvement in the quality of the labor force. A policy which increases the capital-labor ratio (for example, by increasing the rate of investment) can lead to permanently higher *levels* of income, but only a temporarily higher growth *rate* (unless it also leads to an increase in the rate of technical change). Here's why: real per capita income grows at the rate of technical change and labor quality improvement, given some capital-labor ratio. Now along comes a policy (perhaps tax policy) that increases the desired capital stock of firms (or perhaps more accurately, the desired wealth of the population, relative to levels of income). This policy leads to an investment boom for some years which will cause a spurt in the short-run growth rate as the economy moves to a new long-run growth path. Notice that the level of per capita income is permanently higher, but once the transition to the new growth path is complete, the rate of economic growth returns to the original rate given by the underlying factors of the rate of technical change and improvement in labor force quality. This basic conclusion of simple neoclassical growth theory is highly stylized, but it is now standard issue. However, the distinction between the level and rate of growth of income is often confused in policy debates.

It is important to note some potential interactions between the investment rate and the rate of technical change.[6] Learning-by-doing effects in investment may positively link the rates of investment and technical change. A society with a higher investment rate might not only have a temporarily higher growth rate in its transition to a higher growth path, but actually might also increase the long-run rate of growth. In the process of investment, people learn new production processes and discover new potential products. Thus, the rate of investment may positively feed back on the rate of technical change. At the micro level, consider the options opening up in the course of a major project: for example, oil exploration in the frozen tundra or the ocean depths, or space exploration. Just as new technologies arise—sometimes—to meet such challenges, the rate of technical advance may depend on the level of investment, and conversely.[7]

---

[5] Trends in multifactor productivity growth are quite similar to the trends of labor productivity growth (the latter include increases due to a higher capital/labor ratio). If one focuses on what determines the rate of R&D expenditures and innovation, clearly there is some exogenous component. There is a demand for new technology that reflects the obsolescence and cycles of old products and production processes; technologies can mature, and while awaiting major breakthroughs a productivity growth slowdown may occur in the sector. However one divides up the credit for the productivity turnaround in the 1980s, the turnaround itself is not inconsequential.

[6] This and related issues are discussed in many of the papers in R. Landau and N. Rosenberg (1986).

Will these effects be temporary or permanent? They may just take production to best practice more rapidly. But they may also expand the horizons of best practice. Since it is impossible to know what it is possible to know, I am content to consider the potential ultimate knowledge and engineering barriers as sufficiently distant to ignore for the sake of this argument.

While analytical and empirical research on the importance of learning and embodiment continues, I tentatively conclude that the investment rate does indeed affect the long term growth rate, not just the temporary (if lengthy) transition to a higher level of output with the same underlying growth rate. The empirical studies of Jorgenson (this issue), Kormendi and Meguire (1985) and Romer (1987) lend strong credence to this conclusion; otherwise, differences in rates of capital formation could be expected to have small, temporary effects on the growth rate and modest effects on the level of output. I suspect a primary source of this effect is the learning or other external economies generated in specific sectors of the economy, which leads to an aggregate increasing returns phenomenon.[8]

The policies which affect the generation of new technology and therefore long-term growth most directly probably occur one layer down from aggregate GNP, in the composition of spending in the economy. Government spending on research and development and physical and human investment can affect the rate of technical change and, of course, spending on human investment such as education may increase productivity directly. Probably the most important of these categories is direct government support of research and development. Table 2 presents some data on recent trends in federal government support of R&D and physical investment expenditures. As can be seen from the table, real government R&D spending and real government physical investment are both substantially lower as a percentage of GNP than in the 1960s, although defense investment and R&D have made something of a comeback in the last few years. (See Boskin, Robinson and Huber (1988) for estimates of government capital formation).

Tax policies make up a second set of policies that, by affecting the way in which the private sector utilizes its resources, can affect how new technology is generated. The structure of the tax system can substantially affect the rate of investment in the economy and the rate of R&D spending by changing the costs of such spending relative to their returns. The original 1981 tax reform was clearly pro-investment and pro-research and development, with the research and development tax credit and

[7] Embodiment of technical change may also do so, although I consider this potentially less important. The embodiment hypothesis entails the notion that it is much too expensive to embody new technology in old capital by converting it, and therefore that the rate at which new technology augments productivity depends upon the rate at which new capital is generated: that is, it depends on the investment rate. If technology is other than Cobb-Douglas, embodiment, like learning, creates a positive correlation between the investment and growth rates.

[8] Note that virtually all econometric growth studies assume constant returns; Denison's original famous study of long-run U.S. growth assumed (never estimated) slight increasing returns to scale; but that no single national time series study could possibly hope to identify separate scale and technical change effects.

Table 2

Federal Outlays By Category, Selected Years

(billions of constant 1972 dollars; percentages of GNP in parentheses)

Fiscal	Physical Investment			R&D	
Year	Defense	Non-Defense	Grants-in-Aid	Defense	Non-Defense
1949	4.0	1.8	0.9	2.0	0.5
	(0.8)	(0.4)	(0.2)	(0.4)	(0.1)
1958	25.2	1.9	2.8	4.7	1.4
	(3.7)	(0.3)	(0.5)	(0.7)	(0.2)
1968	30.2	3.4	7.6	11.4	10.2
	(2.9)	(0.3)	(0.7)	(1.1)	(1.0)
1978	15.8	4.4	11.0	7.6	7.9
	(1.1)	(0.3)	(0.8)	(0.5)	(0.5)
1985	31.1	5.0	9.8	11.6	6.5
	(1.9)	(0.3)	(0.6)	(0.7)	(0.4)

Sources: Office of Management and Budget, Budget of the United States, selected years, and appendices.

accelerated depreciation. The 1986 tax reform extended the research and development tax credit for three years, but reduced it to 20 percent while tightening eligibility. However, other investment incentives were severely restricted, and the reduction in tax rates only partly compensates for this tightening. The net effect is somewhat anti-investment.[9] Obviously, overall fiscal and monetary policy may also affect the allocation of resources to investment and R&D, and hence the level and perhaps growth rate of output.

Tax Reforms In The 1980s and Their Economic Effects

The two major tax reforms—those enacted in 1981 and 1986—were promoted as beneficial to economic growth. A substantial fraction of the time of many professional economists in the 1980s was spent analyzing the (likely) effects of these reforms on the economy. As noted above, the impact of a tax policy on economic growth over any particular period of time may stem from several sources. For example, lower and more equal marginal tax rates might improve resource allocation, shifting capital and/or labor from lower to higher productivity uses, gross of tax. While this effect would cause only a one-time shift to a higher level of output, and therefore increase the growth *rate* only temporarily, the effect potentially could be large and take substantial time to occur. Particular features of the tax law could alter the net of tax cost and/or returns from R&D spending and, perhaps with a few years lag, affect productivity growth. Investment incentives might affect capital formation and temporarily alter the growth rate, or even affect it permanently if the altered rate of capital formation changed the rate of technical change. Indeed, a reform could have several offsetting effects. Removing the investment tax credit *might* lead to more efficient use of the existing capital stock, and temporarily boost productivity; simultaneously, it *might* decrease the fraction of GNP devoted to capital formation, which could offset or overwhelm the atemporal efficiency effect.

---

[9] Large deficits also ultimately crowd out private investment and/or crowd-in foreign capital, as discussed below. We must also ask whether the increased borrowing is financing government investment or research and development or whether it is simply financing consumption. Monetary policy also can affect the composition of output and the level of aggregate demand in the economy, since it potentially affects the before-tax cost of capital (by affecting interest rates and their term structure), and thus the real cost of embarking on R&D and investment projects.

The balance of this paper discusses the major features of the different tax reforms and evaluates their likely economic effects. Because public finance economists devote so much attention to efficient use of the capital stock, an evaluation of the potential (if one-time) efficiency gains attributable to the reforms is a necessary complement to the potential long-run and possibly permanent effects on the rate of growth.

The central features[10] of the 1981 Economic Recovery and Tax Act (ERTA) and 1982 Tax Equity and Fiscal Responsibility Act (TEFRA) were to: 1) Phase in reduction of marginal personal rates (including bracket indexing by 1985); 2) Reduce the top-bracket rate of 70 percent on investment income to 50 percent; 3) Adopt the Accelerated Cost Recovery System (ACRS) and extend the investment tax credit (ITC); 4) Adopt universal Individual Retirement Accounts (IRAs); 5) Implement the incremental Research and Development Tax Credit.

The tax cuts embodied in ERTA were projected to reduce revenues by some $787 billion over the period between fiscal year 1982 and fiscal year 1988. Over $500 billion of this total was a reduction in individual income taxes, $200 billion was in corporate income taxes, and $40 billion in other taxes. In 1982 TEFRA reduced this total tax cut by $244 billion: $72 billion less in individual reductions, $131 billion less in corporate reductions, and $41 billion less in others. While these revenue effects could have substantial impacts on the economy in and of themselves, I focus here on the structure of tax policy—that is, on changes in the tax structure to lower rates for individuals, accelerated depreciation, and other structural changes—without, for now, considering amounts collected.

The first thing to notice is that ERTA/TEFRA lowered effective marginal rates on capital income and reduced the substantial variation in marginal tax rates on alternative investments (see Table 3). The combination of high and fluctuating

---

[10] I focus here on growth and investment efficiency. There are, and were, other goals of tax reform, such as simplification, fairness, and administrative cost reductions. Some progress has been made in some of these dimensions, but what has been accomplished in these dimensions is more in the eyes of the beholder than in any serious analysis. For example, the tougher alternative minimum tax provisions are designed to increase —by some definitions—fairness, but clearly make the tax laws more complex and raise costs when taxpayers have to keep track of two sets of tax laws. Vickeryesque notions of averaging for fairness and efficiency were dumped in favor of revenue. Even limited income averaging was abolished—and listed as simplification. It takes some bravado to claim credit for simplifying the tax laws for a subset of taxpayers who would prefer them to be more complex. The lower marginal rates, of course, make up for some of these problems, by making the government a smaller silent partner in many decisions.

Table 3

Effective Tax Rates By Asset Type[a] (selected major categories)

Asset	*pre-1981*	*ERTA*	*ERTA / TERFA*
Equipment			
Autos	17.0	−32.8	9.6
Office, Computing, and Accounting Equipment	2.3	−49.4	11.9
Trucks, Buses, and Trailers	10.1	−45.2	11.3
Service Industry Machinery	20.3	−28.5	8.3
Electric Transmission and Distribution Equipment	29.2	3.2	24.2
Engines, Turbines	31.8	16.3	30.2
All Equipment	17.2	−18.8	11.4
Structures			
Industrial Structures	49.6	38.4	38.4
Commercial Structures	46.8	35.6	35.6
Farm Structures	41.1	35.8	35.8
All Structures	40.8	30.0	36.1

[a]*Source:* Jane Gravelle, "Capital Income Taxation and Efficiency in the Allocation of Investment," *National Tax Journal*, 1983.

inflation and the unindexed (especially corporate) tax system increased the level and variation in tax rates on capital income in the corporate sector (Feldstein and Summers, 1979). It was widely, but not universally, believed that this both retarded capital formation in the 1970s and misallocated the capital stock. The studies assessing the deadweight loss from misallocation of the capital stock generally conclude that ERTA/TEFRA substantially *improved* resource allocation—that is, reduced economic distortions—relative to the pre-1981 law. For example, Gravelle (1983) estimates that the deadweight loss from nonneutral taxation of corporate capital assets was only $1.5 billion in 1980 dollars, compared to a loss of $3.3 billion (in 1980 dollars) under pre-1981 law. She only examines the misallocation of the capital stock among different uses and does not deal at all with the distortions over time—that is, in decisions to spend or save—which, after all, were the primary motivation for accelerating depreciation schedules, expanding the investment tax credit, and so on. Auerbach (1983) also estimates that the welfare cost of the distortions within and between industries was both small and approximately half under ERTA/TEFRA of the average over the previous decade. Fullerton and Henderson (1984) also estimate that the tax-induced distortion in the allocation of capital was surprisingly small. Some potential problems with the typical effective tax rate calculations underlying such studies are discussed below. Note also that these static general equilibrium model calculations usually ignore costs of adjustment, assuming capital is malleable and rapid equalization of rates of return, net of tax. Such adjustment costs may well be nontrivial. Apparently, for example, a large percentage of commercial real estate was temporarily unemployed following the partially tax-induced construction boom.

While neutrality among types of investment is desirable,[11] measuring economic depreciation may be quite difficult. Most studies have assumed a common corporate discount rate across types of assets and among industries. This assumes either the absence of risk or that all asset- or industry-specific risks can be diversified away in portfolio decisions.

A second conceptual problem is whether all types of assets, and investments in all industries, have equal access to debt as opposed to equity finance. The ability to borrow, deduct the interest, and still claim depreciation allowances reduces tax rates on leveraged investments. Do allegedly heavily taxed structures have greater access to debt finance than allegedly subsidized equipment, or is there a systematic relationship between the structure or equipment intensity of production and the ability to borrow? Boskin (1985a) and Gordon, Hines and Summers (1986) question this basis for the types of effective marginal tax rate calculation discussed above and conclude that the assertions that ERTA/TEFRA heavily taxed structures while subsidizing equipment may have been dramatically overstated. Stiglitz (1976) demonstrates how the corporate tax may affect entrepreneurial activity and risk-taking, especially when there are credit constraints. Further, advertising and goodwill expenses are written off in the first year, whereas capital outlays for plant and equipment are amortized over longer periods of time. The same is true of firm-specific human capital expenditures such as training programs and managerial reorganization.

Given that owner-occupied housing (accounting for perhaps one-third of all tangible capital) is not taxed in the United States, slowing the rate of depreciation and eliminating the investment tax credit *may* result in some efficiency gains among types of plant and equipment, but *worsen* the distortion between housing and tangible business investment.

Taxes may also affect human investment, but the major feature of most human investment is that it is expensed: for example, paid for out of foregone earnings, which are untaxed (see Boskin (1975) for the analytics and optimality criteria for taxation of human and tangible investment).[12] Thus, human investment is somewhat favored over ordinary investment (to the extent it can be financed similarly). The slowing of depreciation for ordinary investment worsens this distortion, although lower rates reduce any advantage ordinary investment might have in obtaining debt finance. Recall that an increased rate of human investment operates like increased Harrod-neutral technical change in raising the growth rate, whereas an increase in the supply of labor at given levels of human investment will raise the growth rate temporarily, but the level of output (ignoring the opportunity cost of leisure) permanently.

[11] More precisely, equal taxation of capital in all cases is desirable unless there are substantial pure profits which cannot be fully taxed (perhaps for administrative or identification reasons).

[12] In recent years, the direct cost of education—tuition—has apparently risen as a fraction of the total cost but still remains well under one-half of the total cost.

> Lesson 2: Given the modest tax-induced distortions in resource allocation, the efficiency gains from equalizing effective marginal tax rates on alternative investments is small. Further, measuring effective marginal tax rates on alternative assets may be quite difficult and achieving equal effective marginal tax rates is probably impossible except in a (pure) consumption tax. Also, attempts to equalize effective marginal tax rates among alternative nonresidential ordinary investments may worsen distortions between such investments and owner-occupied housing and/or human investment.

A fundamental result of recent research in public finance is that the only neutral tax treatment of investment (ignoring other decisions such as labor supply, which taxes may distort) is any combination of depreciation allowances and interest deductibility which yields the same present value of deductions as expensing (Atkinson and Stiglitz, 1980). Simply put, equity financing should be expensed, as should debt-financed investment if borrowing is brought into the tax base. Any tax system which allows slower depreciation than this is discriminating against investment in favor of consumption.[13]

Various attempts to measure intertemporal distortions generally conclude that they are many times larger than distortions among sectors or assets. For example, Boskin (1978) estimates the annual welfare cost of the tax-induced intertemporal distortion of the timing of consumption to be about $50 billion in 1976, about 2 or 3 percent of national income reflecting the pre-ERTA/TEFRA tax rules. Fullerton, Shoven and Whalley (1983) use a disaggregated general equilibrium model to estimate the efficiency gains from switching from pre-ERTA/TEFRA rules to a progressive consumption tax at about 1 to 3 percent of national income, the larger figure reflecting simultaneously integrating the corporate and personal tax and "upper range" parameter assumptions. In their calculations, moving to a pure income tax would *decrease* welfare unless the corporate and personal taxes were integrated, in which case it would break even relative to (then) current law. Jorgenson and Yun (1986) estimate that moving to a consumption tax and raising labor income taxes to compensate for lost capital income tax revenue would raise the *present value* of welfare by $1.9 trillion in 1980, about 3 percent of national income annually once annuitized.[14] The move to an income tax with equal tax rates on all assets but with the intertemporal distortions remaining would achieve only one-third of this benefit.

---

[13] Risk-taking and asymmetries result in important addenda to this rule. For examples, see Gordon (1985) and Auerbach and Poterba (1986).

[14] We assume the economy is below the golden rule capital intensity; the usual statement in a riskless world is that the marginal product of capital exceeds the growth rate. In a risky world, such simple comparisons are inappropriate and in any event, the expected returns to risky capital has exceeded the growth rate, which in turn has exceeded the safe rate (like the real rate on government securities). In a risky world, the nature of the source of the risk is one of the factors to consider in deciding optimal saving rules. Again, I do not believe that the United States is in danger of oversaving.

Obviously, estimates of the intertemporal welfare loss are based on estimates of the long-run response of saving to the after-tax real return. The estimate of the interest elasticity of saving of 0.4 presented in Boskin (1978) forms the base estimate in many of these studies (for example, Fullerton, Shoven and Whalley, 1983). Jorgenson and Yun (1986) assume maximization of the discounted sum of utilities, which implies the capital stock is perfectly elastic in the long-run. While I believe a modest positive interest elasticity is the best point estimate, considerable uncertainty still surrounds the estimates. Summers (1984) presents evidence of elasticities of private saving with respect to the after-tax real rate of return larger than that originally estimated by Boskin. Several other researchers have found very small elasticities. Boskin and Lau (1988) stress the importance of disaggregating by wealth, age, and vintage in evaluating interest elasticities.

Estimates of the size of the tax wedge between returns to savers and on investment is a matter of some controversy. First, the intertemporal distortion induced by the corporate tax depends upon the marginal source of finance; most studies assume the marginal and average debt and equity contributions to marginal finance are the same (one-third and two-thirds, respectively). If tax-preferred debt finance is the sole source at the margin, then the corporate tax may just be an inframarginal non-distortionary tax on shareholders (Stiglitz, 1976). On the personal side, several forms of saving benefit from various tax preferences: housing, durables, employer pensions, and so on. What fraction of savers and saving is taxed on the return at the margin is open to question. The personal tax has always been a hybrid of a consumption tax and an income tax, moving closer to one or the other base as certain features (for example, universal IRAs) were changed. My view is that, at both the corporate and personal income tax levels, substantial intertemporal distortions occur. Despite tax advantages, I do not believe all marginal finance is debt, at least for many firms and industries, and I believe most households save—partly for liquidity reasons—in taxable assets at the margin. The overall size of this tax wedge, however, is much less than simple textbook comparisons of income and consumption taxes indicate due to the factors mentioned above.

Even if the tax system did not distort saving and investment choices by driving a wedge between the before and after tax rate of return, it could still do so by raising the average tax rate on income from capital. If credit constraints (perhaps due to imperfections in information) lower the relative cost of internal finance, corporate cash flow will be an important determinant of investment (Stiglitz and Weiss, 1981), a finding common in large macroeconometric models.

Since ERTA/TEFRA substantially reduced the cost of capital and the average effective marginal tax rate on new investment (see the estimates in Table 4), ERTA/TEFRA was pro-investment. The two bills substantially redressed the misallocation caused by our corporate and personal income tax system toward consumption away from saving and investment. They certainly did not do so perfectly, but relative to pre-1981 law, the worst distortions, the intertemporal ones, were reduced considerably.

Table 4
Effective Tax Rates for All Assets, 1965–1982, Selected Years

	Auerbach[a]	*Hulten-Robertson*[b]
1965	35.7	26.3
1970	49.7	52.3
1975	37.0	32.1
1980	31.9	33.1
1981	17.7	4.7
1982	24.6	15.8

[a] Alan Auerbach, "Corporate Taxation in the United States," *Brookings Papers on Economic Activity,* 2:1983. Auerbach assumes a 4% real after-tax rate of return and forecasts inflation based on past values.
[b] C. Hulten and J. Robertson, "Corporate Tax Policy and Economic Growth," Urban Institute Discussion Paper, December, 1982. These estimates assume a 4% real after-tax rate of reurn and, for 1981 and 1982, a 6% rate of inflation.

> Lesson 3: The gains from removing intertemporal tax distortions—for example, by moving to pure consumption taxation—are potentially large.

Thus, I conclude that tax rules affect both intertemporal and intersectoral resource allocation, but that levelling the playing field among types of assets, while desirable, is quantitatively less important than levelling it between consumption and saving. While (under certain separability conditions) both could be accomplished with a pure consumption tax,[15] constraints on the taxation of owner-occupied housing apparently leave the economy with only the possibility of a second-best optimality. Even if the evidence is currently debatable, the *potential* efficiency cost and at least temporary productivity growth effects from intertemporal distortions are surely greater than those from the recent misallocations of existing capital.

Likely Effects of the 1986 Tax Reforms

The most important features of the 1986 tax reform were: 1) Lowering personal and corporate income tax rates (to a maximum of 28 percent and 34 percent, respectively); 2) A substantial shift (amounting to about $120 billion over the next five years) of the tax burden from the individual to the corporate tax; 3) Elimination of the investment tax credit (a feature common to all the major reform proposals); 4) Much slower depreciation schedules; 5) Stiff alternative minimum tax for corporations

---

[15] A pure consumption tax would expense investment thereby both equating effective tax rates on marginal investments of different types at zero (but positively taxing inframarginal investment) and leaving intertemporal consumption choices undistorted. True economic depreciation in an income tax would accomplish the former goal of atemporal neutrality but not the latter goal of intertemporal neutrality.

(to insure that no corporation that reports current profits to its shareholders will avoid paying taxes);[16] *6) The tax deductibility of Individual Retirement Accounts (IRAs) will be income-tested, and other tax-deferred retirement accounts (like 401(k) plans)* will limit the amounts that individuals can contribute; 7) Capital gains will be taxed in full as ordinary income, with no inflation adjustment for the basis of the asset; 8) Extension of the research and development (R&D) tax credit to December 1988 with tightened eligibility and a reduction in the rate from 25 percent to 20 percent; 9) Elimination of tax deductibility of consumer debt; 10) Various other changes in accounting rules, industry-specific items, and the personal tax base (for example, state and local sales taxes will no longer be deductible, the personal exemption is increased, and income averaging is eliminated).

The 1986 tax reform is extremely complex. A few basic features deserve considerable attention, especially in light of the accrued evidence noted below concerning the effect of the supply-side structural tax cuts of 1981 on investment, and to a lesser extent, of IRAs on saving, and the history and instability of tax reform.

First, it is likely on balance this tax reform is pro-consumption and anti-saving and investment. It is easy to overstate this fact by focusing only on specific incentives, such as the removal of the investment tax credit, slower depreciation, the full taxation of nominal capital gains as ordinary income, or the sharp limits on tax deferred saving. Offsetting these are lower marginal tax rates which should increase saving slightly, and more importantly, sharply reduce the incentive to borrow at both the corporate and personal level. Recall that with interest payments deductible, there is a tax advantage to debt-financed (as opposed to equity-financed) corporate investment and also that consumers have a substantial advantage in debt finance of purchases, such as automobiles. The reduction in the incentive to borrow, because of the much lower marginal tax rate for both corporations and individuals, is augmented by the elimination of the deductibility of consumer interest payments.[17]

The investment tax credit and depreciation schedules are very powerful determinants (in addition to expected general business conditions) of investment. Two reasons for the sharp reduction in investment spending in 1986 were the uncertainty about what tax rules would ultimately prevail and that all the major reform proposals suggested removing the investment tax credit retroactively to January 1, 1986.

[16] A commonly used estimate is that the extra $120 billion projected to be raised from the corporate tax over the next five years is composed of $180 billion in increased taxation on new investment and $60 billion in tax reductions on old capital. This reduction in corporate cash flow may decrease investment due to credit constraints caused by information effects, lowering the relative cost of internal finance. A large fraction of the tax increases are industry-specific, especially those with respect to defense contractors, real estate, and financial institutions. For example, banks will no longer be able to tax arbitrage tax exempt bonds; passive loss rules for real estate tax shelters will be tightened substantially; defense contractors will have to make changes in their accounting; and so on.

[17] At first glance, it might seem that it would be easy for homeowners to circumvent this provision by increasing their mortgage, and using the extra equity to finance desired purchases, such as an automobile, while deducting the extra interest. At least nominally, however, the new tax law confines the use of tax deductible interest payments on mortgages to housing, educational financing, and medical expenditures. Undoubtedly, this rule will be difficult to enforce, but combined with the lower marginal tax rates, it should result in less borrowing on average over the business cycle in the years ahead.

Summers (1987) argues that investment tax credits and depreciation allowances are very important because corporations use extremely high rates of discount for future cash flows. Thus, cash flows in the near future, such as those that occur instantly with the investment tax credit and very quickly with accelerated depreciation, are given disproportionate weight in investment decisions. I believe there is something to this argument and that, other things equal, the net impact of the 1986 tax reform will be to retard capital formation, unless it is offset by macroeconomic policy and conditions.

The 1981 tax cuts were implemented at the beginning of a substantial disinflation. Since it was clear a recession would result, it was unreasonable to expect an investment boom as demand fell, inventories accumulated, and capacity utilization rates plummeted. Some concluded from this that the "supply-side" tax cuts had failed, a conclusion which is no more justified than the counter assertion that the investment boomlet of 1983–84 was due entirely to the tax changes. Nevertheless, the tax changes did play an important role, as the investment boom occurred even in the face of inordinately high real interest rates over the period. Some quantitative estimates are discussed below.

Similarly, corporate saving, which had declined in the late 1970s and remained quite low in 1981 and 1982, rebounded to more typical levels by 1984–85. As shown in Table 5, personal saving declined to postwar lows in 1985–86. Why didn't the tax cuts generate increased private saving? First, IRAs did induce some additional saving. While some of the movement of funds into IRAs at the *outset* was from currently taxed assets rather than from new saving, Venti and Wise (1987) estimate that about one-half was net new saving. As less fungible taxed assets remained to be switched over time, the IRAs would have had a greater impact, as the success of registered retirement accounts in Canada and postal saving accounts in Japan in contributing to higher private saving in these countries suggests. Further, the very high interest rates in the 1980s accrue disproportionately to older people in the population, who because of their natural life-cycle saving pattern, own a disproportionate share of assets (Boskin, Kotlikoff and Knetter, 1985). Many of these people are at a time in their life when they have stopped saving. Therefore, we would not expect an increase in real interest rates to increase their saving, and it might even accelerate their dissaving. As they planned on greater returns from their assets to finance their retirement, they could afford to consume more, thus offsetting the increased saving done by the still working population. Further, both high real interest rates and increased share values led to a sharp decline in contributions to defined benefit pension plans (Bernheim and Shoven, 1986).

The big news in saving in the 1980s, of course, was federal government borrowing. Clearly, the increase in federal borrowing swamps any potential increase in private domestic saving and much of the difference is being met by borrowing from abroad. (Compare the last two columns in Table 5 with the data for 1950–70).

In analyzing the overall impact on capital formation, of course, saving *and* investment play a role. Investment—the demand for new capital goods—was stimulated substantially by the structural changes in tax policy. Saving, at least personal saving, probably was little affected in the short run by the supply-side policies. Had

Table 5
U.S. Net Saving and Investment, 1951–87
(as percent of gross national product)

	1951–60	*1961–70*	*1971–80*	*1981–86*	*1987*
Total Net Saving	7.1	7.8	7.0	3.0	1.8
Net Private Saving	7.5	8.2	8.0	5.9	4.1
Personal Saving	4.7	4.8	5.5	4.1	2.3
Corporate Saving	2.8	3.4	2.4	1.9	1.8
State-Local Govt. Surplus	−0.2	0.1	0.9	1.4	1.2
Federal Govt. Surplus	−0.2	−0.5	−1.8	−4.3	−3.5
Total Net Investment	7.3	7.7	7.0	3.0	1.6
Net Foreign Investment	0.3	0.6	0.3	−1.6	−3.5
Private Domestic Investment	7.0	7.1	6.7	4.6	5.1
Plant and Equipment	3.0	3.5	3.3	2.3	1.6
Residential Construction	3.3	2.6	2.8	1.9	2.6
Inventory Accumulation	0.8	1.0	0.7	0.3	0.9
Memoranda: Capital Consumption	8.9	8.4	9.8	11.3	10 6
Gross Private Saving	16.2	16.6	17.8	17.2	14.7
Real Gross Investment	16.0	16.5	17.2	16.7	17.5

Data are averages of annual flows, as percentages of gross national product. Total net saving and total net investment differ by statistical discrepancy. Detail may not add to totals because of rounding.
Source: U.S. Department of Commerce.

the supply of foreign capital proved less elastic, the lack of increased saving in the face of increased demand for investment would have just raised interest rates. The very elastic supply of saving from the rest of the world allowed an increase in investment (as presaged by Harberger, 1980). Note also that the distinction between the marginal rate of substitution of present for future consumption and the marginal rate of transformation in production is likely to be much larger, at least for the short and medium run, from this international perspective than those estimates discussed above.

Lesson 4: International capital flows break the link between domestic saving and investment rates, at least in the short run.

Whether they do so in the long run remains to be seen; there apparently is no compelling case in economic history of an advanced economy financing long-term growth by importing capital. The pressures created by current account imbalances and large annual negative flows of net foreign investment implying future streams of interest and dividend payments to foreigners and the subsequent need for greater net exports probably create some long-run links between domestic saving and investment rates (Summers, 1986; Feldstein and Horioka, 1980). I still believe that the investment rate in advanced economies is *ultimately* constrained by the supply of domestic savings, and hence tax policies discouraging saving eventually will affect domestic investment.

In any event, it is likely that the impact of the deficits is far greater than those of the structural tax policies.

An important feature of the 1981–82 corporate tax changes is that the extended investment tax credit and accelerated depreciation were available only on *domestic* investment. Therefore, the policies not only had an impact on the demand for capital, but certainly increased investment *in the United States.* Boskin and Gale (1987) document the effect of the investment incentives on the international location of direct investment.

Throughout the Reagan Administration it has been unclear which taxes would remain permanently, and which would be abandoned or altered by Congress. For example, TEFRA eliminated much of the business tax reduction of 1981, the Deficit Reduction Act of 1984 changed various other provisions, and the 1986 tax reform reversed many features adopted in 1981. This record of continuing change cannot avoid producing great uncertainty over the durability of any tax reform. This may have accelerated some investment in response to perceived temporary tax advantages, and slowed other investment.

Major increases in investment affect the capital stock and productivity only over a prolonged period of time. Leaving out cyclical influences, it would take years to increase the capital stock substantially, and therefore, to realize major productivity gains. Each year, net investment is a small fraction of GNP and a smaller fraction of the existing capital stock. Even if the new investment embodies more up-to-date technology, it will take a decade or more for the standard of living to change noticeably because of it.[18]

As Alan Blinder (1984) notes, even phenomenally successful investment incentives will only add a modest amount to the long-run growth rate. Blinder wrote, "[I]f the share of investment in GNP is raised by 3 percentage points for a full decade—and that would be a substantial achievement—this would lead to a one percentage point increase in the growth of the capital stock, leading after ten years to a capital stock 10.5 percent higher. With a share of capital in output of 30%, the level of real GNP would be 3.5% higher. That would be a spectacular achievement. If we take a 25 year perspective, it amounts to adding only 0.12% to the GNP growth rate."

Tax Incentives and Investment

Many studies document the impact of ERTA/TEFRA on the user cost of capital and effective tax rates on a *marginal* investment for the aggregate economy, separately by structures and equipment, disaggregated by type of structure, type of equipment, and by industry. The range of changes in the user cost of capital estimated by Boskin (1985b) and presented in Table 6 is quite similar to those estimated in earlier prominent studies, including Hulten and Robertson (1984), Gravelle (1983, 1985),

---

[18] Summers (1980) estimates that doubling the net investment rate from 4.5 percent to 9 percent would raise the productivity growth rate by only 0.3 percentage points per year (ignoring any embodied technical change) over the first decade, then declining (in the steady state, the growth of output is independent of the investment rate in this model). If all technical change is embodied, the short-run impact approximately doubles.

Table 6

A. Percentage Decrease in User Cost of Capital Due to Investment Incentives in ERTA / TEFRA

User Cost Concept						
R	*Wts.*	*I deflator adj.*	*1981*	*1982*	*1983*	*1984*
(1) Constant	I	-	5.9	5.4	3.0	3.0
(2) Constant	I	+	5.9	5.4	3.0	3.0
(3) Constant	K	–	9.9	8.5	6.0	6.0
(4) Constant	K	+	9.9	8.5	6.0	6.0
(5) FED	I	–	6.5	5.9	2.8	3.0
(6) FED	I	+	6.5	5.9	2.8	3.0
(7) FED	K	–	10.0	8.3	6.3	6.4
(8) FED	K	+	10.0	8.3	6.3	6.4

R is real required rate of return, assumed either constant at 4%, or from FED data in Sahling and Akhtar (1985); wts. refers to how equipment and structures are weighted; I refers to flows of investment in each category, K to capital stocks in each. I deflator adj. indicates whether or not the user cost is adjusted for the decline in the investment goods deflator relative to the GNP deflator.

B. Impact of ERTA / TEFRA on Investment

*Direct calculation of change in desired capital stock, change in net investment assumed spread over 3 or 5 years and change as % of net investment.**

Change in Desired Capital Stock	*Change in Net Investment ($billions) per year*		*If Spread Over 3 yrs. Change per year as % of average net investment 1981–4*
	3 yr.	*5 yr.*	
7.3	29.1	17.5	23.9%

*Assumes unitary elasticity of desired capital stock with respect to the cost of capital and change in user cost is simple average of 1982 and 1983 changes above for capital stock weights cases.

Auerbach (1983), and Fullerton and Henderson (1984). Each study reports substantial decreases in the user cost of capital or in effective tax rates on a marginal investment due to the investment incentives in ERTA/TEFRA. They generally conclude that while ERTA/TEFRA decreased the disparity in user costs or effective tax rates on a marginal investment across industries, substantial variation remained.[19]

---

[19] A very important caveat to those estimates must be mentioned: the definition of structures and equipment used in the national income accounts is quite different from those in the tax laws. For example, much equipment that was treated as having 5-year life under ACRS is considered a structure for national income account purposes since it is "bolted down," like a rolling mill. Therefore, one must be very cautious in attempting to use disaggregated national income account equipment and structure numbers, especially for recent years, to estimate the impact of the tax laws on investment in equipment and structures.

Hulten and Robertson (1984) and Gravelle (1985) also estimate changes in desired capital stocks on the assumption that the elasticity of the desired capital stock with respect to the rate of return is unity. They tend to take this as the standard case when they make such calculations. I have done likewise in Table 6, but the reader should be informed both by the caveats mentioned below, and the earlier debate on this issue between Hall and Jorgenson (1969). and Eisner (1969). The changes in the desired capital stock, as reported in Table 6, on the order of 5 percent or more seems to be roughly in accord with the estimates of other researchers. Spread over 3 or 5 years, this change would lead to a very substantial increase in investment, accounting for about 20 percent of net investment in the 1982--84 period.

The conclusion that the investment tax incentives are important is fairly robust to alternative models of business investment behavior. For example, Summers (1981) uses a *q*-theory approach and Feldstein and Jun (1987) use a return over cost model and conclude *new* business investment is substantially responsive to reductions in its taxation.[20] (Other studies attributing a powerful role to tax incentives for investment include Sahling and Akhtar, 1984; Sinai and Eckstein, 1983; and Makin, 1984.)

Contrary opinions are expressed by Bosworth (1985), and Eisner and Chirinko (1983). Of course, Lucas (1976) reminds us that standard econometric policy evaluation *may* be misleading. Also, many factors *other* than taxes influence investment. These include the price of investment goods (and the price of computing fell rapidly in this period, more than offsetting the possible slight increase in effective tax rates), interest rates, the nature and course of a recovery and expectations about future economic activity.

However, my best estimate is that the investment incentives in ERTA/TEFRA were responsible for at least 20 percent additional net investment in the United States in the 1982–84 period, less in 1985–86. The additional investment, at least temporarily, raised the growth rate and permanently raised the level of potential output. They were probably a modest part of the explanation of the apparent partial productivity rebound in the 1980s.

> Lesson 5: Tax policy is an important (but hardly exclusive) determinant of investment.

I estimate that the Tax Reform Act of 1986 will raise the effective marginal tax rate on new business investment by about 10 percent, with standard assumptions regarding inflation and required rates of return. The user cost of capital will increase by much more on some equipment. The lower tax rates do not fully offset the elimination of the investment tax credit and the slowing of the depreciation. New capital will be taxed more heavily, whereas old capital will be taxed less heavily. If the reform remains roughly in its current form, the U.S. rate of investment will be lower than it would otherwise have been for several years. This in turn should slightly decrease the short-run growth rate, and the level of potential output. Given the vagaries of discounting nominal cash flows, a return to 1979–80 inflation rates will raise the (weighted average) effective tax rate on new investment by substantially more than 10 percent. Some of this deleterious effect on the level of investment and potential output might be offset by greater tax neutrality (although recall from Lesson 2 that these efficiency gains may be quite small).

---

[20] Indeed, Feldstein and Jun estimate that each percentage point increase in the difference between the real net return and the cost of funds increases the ratio of fixed nonresidential investment to GNP by 0.3. The increase induced by ERTA/TEFRA and the decreased inflation of 1981–85 accounts for two-thirds of the increase in net investment, by these estimates. These estimates suggest that for each dollar of (prospective) business tax collections, investment will fall by more than a dollar.

Tax Cuts, Deficits, and Their Impact on GNP and Its Composition

Thus far, I have limited the discussion to various structural aspects of the tax laws, especially effective marginal tax rates on new investment. But tax revenue changes may also affect the level, growth, and composition of GNP. First, how much were taxes reduced in the 1980s, and what fraction of the large federal government budget deficits are due to tax cuts rather than to other factors? Second, what impacts did the deficits have on the level and composition of GNP?

It is commonly thought that the large tax cuts instituted in ERTA and the substantial Reagan defense buildup are the primary causes of the large budget deficits. A corollary is that the president and Congress have been stymied, unable to cut defense spending or to raise taxes and have made only minor cuts in other programs, resorting to accounting gimmicks to reduce the current budget deficit. While these claims have an element of truth, they are far from a full and accurate account.

Projections of the policies in place prior to President Reagan's changes in 1981 indicate that the United States would have had a balanced budget by the late 1980s or sooner. A decomposition of the budget deficit based on these pre-Reagan projections would show large projected tax increases and attribute about two-thirds of the budget deficit to tax cuts, one-third to spending increases. A large part of the spending increase came in Social Security and Medicare, and interest payments (both because interest rates rose and the debt grew rapidly). The tax cut was made much larger by the rapid reduction of inflation in the early 1980s at a time when tax rate brackets were not yet indexed. There have been three major (and some minor) tax increases since then. In 1982, TEFRA rolled back a substantial fraction of ERTA (and disproportionately on the corporate side); the Social Security Amendments of 1983 raised taxes; and the Deficit Reduction Act of 1984 did likewise. Federal taxes as a share of GNP are about 19 percent, about the same level as in 1979 and 1980, and substantially higher than the average for the 1960s and 1970s. While there was rapid growth in defense spending in the early Reagan years, there has been no growth for the last three, and the principal reason the Congressional Budget Office now projects declining rather than rising deficits is the assumption of zero real growth in defense as opposed to a continued defense buildup.

With these facts of fiscal history in mind, it is worth mentioning that there was nothing necessarily optimal about the level, trend, and composition of government spending circa 1980. *Both* political parties had attempted unsuccessfully to curtail the explosion in entitlement programs and both agreed on the need for a defense buildup.

Thus, I believe that a fair summary of the events of the first seven years of the Reagan presidency would be that the original tax cuts, interacting with the disinflation, stabilized tax revenue as a share of GNP at historical levels, but reduced them relative to the projected increase; combined with the growth of Social Security, defense and interest payments, this created the large budget deficits. The major pressure to reduce the budget deficit now comes from the trade deficit, as the budget deficit drains national saving, requiring large inflows of foreign capital and the associated current account deficit. Continued appeals to reduce the role of government or moral imperatives to prevent passing larger debt burdens to future generations appear to have been eclipsed—at least temporarily—in the political economy debates on budget deficits. Further, as discussed by Boskin (forthcoming) and Eisner (1986), the usual deficit data may be quite misleading for various accounting, economic, and conceptual reasons.

First, did the tax cuts stimulate spending, or as in the Barro (1974) intergenerational altruism model, merely call forth increased private saving to pay for the expected future tax liabilities?[21] It is exceedingly difficult to infer from the saving behavior reported in Table 5 (even adjusting these numbers for a durables boom and for capital gains, net of losses, as measured by the FED Flow of Funds data), that national saving increased, let alone enough to offset the budget deficits. Some interesting calculations are presented in Poterba and Summers (1987) indicating that the long-term crowding out effects are dollar for dollar, whereas the short-run stimulative effects of tax cuts are perhaps twenty-five cents on the dollar.

I conclude that the stimulus to aggregate demand is insufficient to raise real incomes enough to offset the budget deficit. I also conclude that, in the United States, the increase in tax revenues for other reasons, such as increased income reporting with lower marginal tax rates, is not sufficient to prevent the budget deficits from occurring. A careful study by Lindsay (1986) indicates that perhaps 40 percent of the revenue lost from the personal tax rate cuts in ERTA was indeed recouped—far from enough to prevent large deficits from occurring, but much larger than the zero effect commonly assumed. In economies with much more advanced welfare states, even a small stimulus may be sufficient to prevent budget deficits from occurring, because in addition to whatever extra revenue is generated because of the extra income, removing unemployed persons from much more generous unemployment benefits and placing them in work, subject to much higher tax rates, will have larger positive net impact on government revenues than in the United States. Blanchard and Summers (1987) suggest this fiscal increasing returns may be prevalent in some European countries.

---

[21] The Ricardian equivalence conjecture suggests that a switch from tax to debt finance should have no impact on national saving: as the budget deficit is exactly matched by increased private saving to offset future tax liabilities. One implication of the conjecture is that the age distribution of resources should not affect aggregate consumption, conditional on the level of aggregate resources. Because deficits are conceptually and practically difficult to measure properly due to inflation, the lack of comprehensiveness of budgets, arbitrary accounting rules, contingent and potential liabilities, and other reasons, an alternative strategy can test Ricardian equivalence without having to measure the current period deficit. Boskin and Kotlikoff (1985) build a finite approximation to an infinitely lived intergenerationally altruistic Barroesque consumption program and test whether given predicted consumption, one can reject the hypothesis that the age distribution of resources does not affect actual aggregate consumption. The hypothesis that the age distribution of resources does not matter is overwhelmingly rejected, thereby rejecting a strong implication of Ricardian equivalence. Numerous other direct and indirect tests of Ricardian equivalence have been developed, and I refer the reader to the excellent discussion by Bernheim (1987). Second, numerous econometric studies have indicated a substantial impact of tax reductions on the current consumption of at least a significant fraction of consumers, far more than would be predicted by forward looking life-cycle savers, let alone Barroesque intergenerational altruists (for example, Hall and Mishkin, 1982; Flavin, 1981).

Thus, I conclude that while far less than the direct effect predicted by myopic Keynesian consumption functions, tax cuts do indeed affect aggregate consumption substantially (Boskin, 1988). Of course, in an open economy close to full employment, the increased spending may drive up interest rates, appreciate the currency and decrease net exports, thereby sharply limiting any stimulative effect on aggregate demand. However, the composition of GNP will shift away from interest sensitive activities such as net exports and investment toward consumption. Surely, the overall fiscal stimulus of the "tax cuts" could not have stimulated employment and income enough to be self-financing.

The impact of deficits on the composition of GNP may be extremely important to long-term growth. From the national income identity, the change in the budget deficit must show up as a change in some component of net saving and investment: net foreign investment, or domestic investment, or private saving, or the state and local surplus. To the extent deficits crowd out investment, they will reduce the growth rate either temporarily or permanently, as discussed above.

Several studies have examined the effects of deficits, variously defined, on the composition of GNP based on postwar U.S. time series data. For example, Boskin (1986), Summers (1986), and Eisner (1986) all conclude from reduced form time series regressions examining the effect of deficits on the composition of GNP that the effects of deficits are substantial. My results suggest that a dollar increase in the budget deficit, given the level of GNP, appears to be associated with a decrease of about 35–40 cents in domestic investment and about 20–30 cents in net foreign investment. Summers (1986) results are qualitatively similar. The results suggest that the link between domestic saving and investment is neither fully nor permanently severed by international capital flows.

These results are relatively robust to alternative specifications and methods of estimation. The primary problem, as with all reduced forms, is the difficulty of inferring causality without a structural model. While these results are from a reduced form of a larger implicit structural model and their explanatory power should not be overstated, they do suggest

> Lesson 6: A shift from tax to debt finance can substantially affect the composition of GNP. These effects are likely to decrease capital formation either in the short run or eventually, with the time pattern depending primarily on the relative interest elasticities of domestic investment and the supply of foreign capital.

This lesson also implies that the conclusion of lesson 4 is only a partial, and perhaps temporary, phenomenon.

Conclusion: Lessons From the 1980s

It is not surprising that following the long-term productivity growth slowdown, the demand for policies designed to reverse the trend has been intense. While the determinants of past growth (which in any event may differ from those of future growth) cannot be precisely quantified and the impact of various tax and deficit policies on these determinants cannot be predicted precisely, some important lessons have still been learned.

My views concerning the effects of tax policy on economic growth, and of recent tax policy and reforms on the productivity slowdown and partial rebound, may be summarized as follows. First, capital formation is quantitatively a quite important determinant of postwar U.S. growth, *at least* as important as the upper range of estimates (see Jorgenson, this issue) from conventional growth accounting studies. The numerous restrictive assumptions in typical growth accounting studies may lead to underestimating the importance of capital formation in determining growth, perhaps even in the long run.

Second, the rate of domestic investment in the U.S. is *ultimately* constrained by the supply of national saving. Thus, tax and deficit policies which directly affect investment or indirectly affect it by altering national saving affect the growth rate over substantial periods, and perhaps even in long-run steady state. The combination of high and fluctuating inflation and the unindexed (especially corporate) tax system in the 1970s raised real effective tax rates on capital income, which in turn contributed to a modest fall in net productive tangible business investment. This in turn was one of the contributing factors to the productivity growth slowdown. The combination of lower inflation and the investment incentives in ERTA/TEFRA lowered effective tax rates on new investment, which, ceteris paribus, stimulated investment, and again was only one of several contributing factors to the partial productivity rebound. The 1986 tax reforms contain numerous offsetting features but, on balance, are likely to retard capital formation, and slow productivity growth slightly relative to what it might have been.

Third, the large deficits of the 1980s reduced national saving substantially. Domestic investment was adversely affected, cyclically adjusted, only modestly because of the large supply of foreign capital which flowed into the U.S. Eventually, that source of saving will have to be replaced by increased national saving or investment will decline substantially.

Fourth, resource allocation improvements may temporarily play a role in improving productivity growth. The worsening distortions in the tax code/inflation interaction in the 1970s (e.g. between housing and tangible business investment) may have played a minor role in the productivity slowdown. The slightly improved atemporal resource allocation apparently achieved in ERTA/TEFRA was too small to have made a quantifiable impact on productivity growth in the 1980s. The Tax Reform Act of 1986 may improve the allocation of resources among types of business investment (that was one of its motivations, but how well it was achieved is unclear), but this effect is likely to be small. A potential exception arises from curtailing abusive tax shelters which could reduce the incidence of unemployed capital resources. In any event, these effects are unlikely to be as important as the capital formation effects.

Fifth, the effects of the tax system on the rate of human investment are difficult to quantify, and are probably less important than the level and composition of government outlays on (or in support of) human investment, and credit conditions for undertaking human investment. While higher education is far from the only human investment, the likely effects of tax policy changes on private higher education investment decisions are complex and offsetting. The higher marginal tax rates on the bulk of the population in the 1970s perhaps caused a slight disincentive for such investment as they raised the cost of the failure of the tax system to depreciate direct expenditures on such investment. The lower marginal tax rates of the 1980s reforms are likely to have a small effect in the opposite direction. But the ability of families to save tax-free to finance such expenditures was first expanded under ERTA/TEFRA (for example, universal IRAs) and then curtailed in the 1986 reform.

Sixth, the effects of the tax system on innovation, which I believe to be an important determinant of productivity growth, are also complex. The incremental R&D credit, especially as tightened in 1986, probably is one reason for the slight R&D spending turnaround (although how much of this is the elasticity of accounting conventions and measurement is impossible to say). This should eventually lead to a slight productivity growth increase. But the demand for the products and processes produced as a result of innovation, like new capital equipment, also has been affected substantially by tax law changes. And surely some of the supply of product and process innovation is a response to this derived demand. As noted throughout this paper, I tentatively conclude that the rate of investment and technical change are positively linked in the long run.

Finally, with due respect to all the caveats and disagreements concerning the size and nature of the effects of tax and deficit policies on the determinants of economic growth as well as the relative importance of each of those determinants, I conclude that tax and deficit policy can be designed to provide an environment that is less inhibiting to growth. Surely many other factors will affect the determinants of the long-term growth performance of the economy. But a tax policy which is more neutral with respect to the consumption versus saving/investment choice, and a deficit policy which does not systematically borrow from the future and from abroad to finance current consumption could combine well to remove some of the impediments to increasing the rate of capital formation, which would raise the short-run growth rate, the level of output, and, to the extent that it feeds back positively on the rate of technical change, the long-run rate of growth.

■ *I am grateful to the Center for Economic Policy Research at Stanford University for financial support and to Lawrence Summers, Joseph Stiglitz and Charles Hulten for valuable advice and suggestions. Timothy Taylor provided both superb editorial assistance and numerous substantive comments.*

References

Abramovitz, M., "Resource and Output Trend in the United States Since 1870," *American Economic Review*, 1956, *46*, 5–23.

Arthur, Brian, "Positive Feedback Mechanisms in Economics," unpublished mimeo, Stanford University, 1987.

Atkinson, Anthony, and Joseph E. Stiglitz, *Lectures in Public Economic*. New York: McGraw-Hill, 1980.

Auerbach, Alan, "Corporate Taxation in the United States," *Brookings Papers on Economic Activity*, 1983, *2*, 451–513.

Auerbach, Alan, and James Poterba, "Tax Loss Carry Forwards and Corporate Tax Incentives," NBER Working Paper No. 1879, 1986.

Baily, Martin, and Alok Chakrabarti, "Innovation, and U.S. Competitiveness," *Brookings Review*, Fall 1985, 14–21.

Barro, R., "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *82*: 1095–1117.

Bernheim, Douglas, "Ricardian Equivalence: An Evaluation of Theory and Evidence," NBER Macroeconomics Annual, No. 2, 1987, 263–304.

Bernheim, Douglas, and John Shoven, "Pension Funding and Saving," NBER Working No. 1622, 1986.

Blanchard, O., and L. Summers, "Fiscal Increasing Returns, Hysteresis, Real Wage and Unemployment," *European Economic Review*, April 1987, *31*, 543–549.

Blinder, Alan, "Reaganomics and Growth." In Hulten, C., and I. Sawhill, eds., *The Legacy of Reaganomics*. Washington, D.C.: Urban Institute Press, 1984, 199–227.

Boskin, Michael, "Notes on the Tax Treatment of Human Capital," *Symposium on Tax Research*, U.S. Treasury Dept., Office of Tax Analysis. Washington: U.S.G.P.O., 1975, pp. 185–195.

Boskin, Michael, "Taxation, Saving and the Rate of Interest," *Journal of Political Economy*, April 1978, *86*,. S3–S27.

Boskin, Michael, "Federal Government Deficits: Some Myths and Realities," *American Economic Review*, May 1982, *72*, 296–303.

Boskin, Michael, "A Closer look at Tax Neutrality Toward Investment," *Tax Notes*, November 11, 1985a, 652–654.

Boskin, Michael, "The 1981–82 Tax Reforms and Investment," mimeo, 1985b.

Boskin, Michael, "Consumption, Saving and Fiscal Policy," *American Economic Review*, May 1988, *78*, No. 2, 401–407.

Boskin, Michael, *The Real Federal Budget*. Cambridge: Harvard University Press, forthcoming.

Boskin, Michael, and William Gale, "New Results on the Effects of Tax Policy on the International Location of Investment." In Feldstein, Martin, ed., *The Effects of Taxation on Capital Accumulation*. Chicago: University of Chicago Press, 1987, pp. 201–219.

Boskin, Michael, Marc Robinson, and Alan Huber, "Government Saving, Capital Formation and Wealth in the United States, 1945–85," In Lipsey, R., and H. Tice, eds., *The Measurement of Saving Investment and Wealth*, Chicago: University of Chicago Press, forthcoming 1988.

Boskin, Michael, and Lawrence Kotlikoff, "Public Debt and U.S. Saving: A New Test of the Neutrality Hypothesis." In Brunner, K., and A. Meltzer, eds., *The New Monetary Economics, Fiscal Issues and Unemployment*. Amsterdam: North-Holland, Carnegie-Rochester Conference Series, 1985, pp. 55–86.

Boskin, Michael, Lawrence Kotlikoff, and Michael Knetter, "Changes in the Age Distribution of Income in the United States, 1968–84," NBER Working No. 1766, 1985.

Boskin, Michael, and Lawrence Lau, "An Analysis of Postwar U.S. Consumption," NBER Working Papers #2505 and #2506, 1988.

Bosworth, Barry, "Taxes and the Investment Recovery," *Brookings Papers on Economic Activity*, 1985, *1*, 1–46.

Denison, Edward F., *Accounting for United States Economic Growth, 1929–1969*. Washington, DC: The Brookings Institution, 1974.

Denison, Edward F., *Accounting for Slower Economic Growth: The United States in the 1970s*. Washington, DC: The Brookings Institution, 1979.

Eisner, Rober, "Tax Policy and Investment Behavior: Comment," *American Economic Review*, May 1969, *59*, 50–64.

Eisner, Robert, *How Real is the Federal Deficit?* New York: The Free Press, 1986.

Eisner, Robert, and Robert Chirinko, "Tax Policy and Investment in Major Macroeconometric Models," *Journal of Public Economics*, 1983, *20*, No. 2, 139–166.

Feldstein, Martin, and Charles Horioka, "Domestic Saving and International Capital Flows," *The Economic Journal*, June 1980, *90*, 314–329.

Feldstein, Martin, and Joo Sung Jun, "The Effects of Tax Rules on Nonresidential Fixed Investment: Some Preliminary Evidence from the 1980s." In Feldstein, Martin, ed., *The Effects of Taxation on Capital Accumulation*. Chicago: University of Chicago Press, 1987, pp. 101–156.

Feldstein, Martin and Lawrence Summers, "Inflation and the Taxation of Capital Income in the Corporate Sector," *National Tax Journal*, 1979, *32*, No. 4, 445–470.

Flavin, Marjorie R., "The Adjustment of Consumption to Changing Expectations About Future Income," *Journal of Political Economy*, 1981, *89*, 974–1009.

Fullerton, Don, and Yolanda Henderson, "Incentive Effects of Taxes on Income from Capital: Alternative Policies in the 1980's." In Hulten, C., and I. Sawhill, eds., *The Legacy of Reaganomics*. Washington, D.C.: Urban Institute Press, 1984, pp. 49–89.

Fullerton, Don, John Shoven, and John Whalley, "Replacing the U.S. Income Tax with a Progressive Consumption Tax: A Sequenced General Equilibrium Approach," *Journal of Public Economics*, 1983, *20*, No. 1, 3–24.

Gordon, Roger, "Taxation of Corporate Capital Income: Tax Revenues Versus Tax Distortions," *Quarterly Journal of Economics*, 1985, *6*, No. 1, 1–28.

Gordon, Roger H., James R. Hines, Jr., and Lawrence H. Summers, "Notes on the Tax Treatment of Structures." In Feldstein, Martin, ed., *The Effects of Taxation on Capital Accumulation*. Chicago: University of Chicago Press, 1987, pp. 223–254.

Gravelle, Jane, "Capital Income Taxation and Efficiency in the Allocation of Investment," *National Tax Journal*, 1983, *36*, 297–306.

Gravelle, Jane, "The Treasury Tax Proposals and Desired Capital Stocks," unpublished mimeo, 1985.

Griliches, Zvi., "Returns to Research and Development Expenditures in the Private Sector." In Kendrick, J., and B. Vaccara, eds., *New Developments in Productivity Measurements*, NBER, *Studies in Income and Wealth*, Chicago: University of Chicago Press, 1980, *44*, pp. 419–454.

Hall, Robert, and Dale Jorgenson, "Tax Policy and Investment Behavior: Reply and Further Results," *American Economic Review*, June 1969, *59*, 388–401.

Hall, Robert E., and Federic S. Mishkin, "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, May 1982, *50*, No. 2, 461–481.

Harberger, A., "Vignettes on World Capital Markets," *Papers and Proceedings of the 92nd Annual Meeting of the AEA, American Economic Review*, May 1980, *70*, No. 2, 331–339.

Hulten, Charles, and James W. Robertson, "The Taxation of High Technology Industries," *National Tax Journal*, September 1984, *37*, 327–345.

Jorgenson, Dale, "Productivity and Economic Growth," Harvard Institute of Economic Research Discussion Paper Series, 1986.

Jorgenson, Dale, "Productivity and Economic Growth," (essay in this volume of *Journal of Economic Perspectives*)

Jorgenson, Dale, and Kun Young Yun, "Tax Policy and Capital Allocation," *Scandinavian Journal of Economics*, 1986, *88*, No. 2, 355–377.

Kendrick, John, "Recent Productivity Trends," *AEI Economist*, August 1986.

Kormendi, R., and P. Meguire, "Macroeconomic Determinants of Growth: Cross-Country Evidence," *Journal of Monetary Economics*, 1985, *16*, 141–163.

Landau, R. and N. Rosenberg, eds., *The Positive-Sum Strategy*. Washington: National Academy Press, 1986.

Lindsey, L., "Individual Taxpayer Response to Tax Cuts 1982–84 With Implications for the Revenue Maximizing Tax Rate," NBER Working Paper No. 2069, 1986.

Lucas, Robert, "Econometric Policy Evaluation: A Critique." In Brunner, K., and A. Meltzer, eds., *The Phillips Curve and Labor Markets, Vol. 1 of the Carnegie-Rochester Conference in Public Policy*. Amsterdam: North-Holland, 1976, pp. 19–46.

Makin, John, "Debt Accumulation and Capital Formation," AEI Discussion Paper No. 1, 1984.

Poterba, J., and L. Summers, "Recent U.S. Evidence on Budget Deficits and National Savings," NBER Working Paper 2144, 1987.

Romer, Paul, "Crazy Explanations For The Productivity Slowdown," NBER *Macroeconomics Annual*, 1987, *2*, 163–202.

Sahling, Leonard, and M. A. Akhtar, "What Is Behind the Capital Spending Boom?" *Quarterly Journal of the Federal Reserve Bank of New York*, Winter 1984–85, 19–30.

Sinai, A., and O. Eckstein, "Tax Policy and Business Fixed Investment Revisited," *Journal of Economic Behavior and Organization*, 1983, *4*, 131–160.

Solow, R., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, 1957, *39*, 312–320.

Stiglitz, J., "The Corporation Tax," *Journal of Public Economics*, April–May 1976, *5*, No. 3, 4, 303–311.

Stiglitz, J. and A. Weiss, "Credit Rationing in Markets With Imperfect Information," *American Economic Review*, June 1981, *71*, No. 3, 393–410.

Summers, L., "Tax Policy and Corporate Investment," NBER Working Paper No. 605, Dec. 1980.

Summers, L., "Taxation and Corporate Investment: A Q-Theory Approach," *Brookings Papers on Economic Activity*, 1981, *1*, 67–140.

Summers, L., "The After Tax Rate of Return

Affects Private Saving," *American Economic Review*, May 1984, *74*, No. 2, 249–253.

Summers, L., "Issues in National Savings Policy." In Adams, G., and S. Wachter, eds., *Savings and Capital Formation: The Policy Issues*. Lexington, MA: D.C. Heath & Co., Lexington Books, 1986, pp. 65–88.

Summers, L., "Corporate Capital Budgeting Practices and the Effects of Tax Policies on Investment." In Feldstein, Martin, ed., *The Effects of Taxation on Capital Accumulation*. Chicago: University of Chicago Press, 1987.

Venti, Steven F., and David Wise, "IRAs and Saving." In Feldstein, Martin, ed., *The Effects of Taxation on Capital Accumulation*. Chicago: University of Chicago Press, 1987, pp. 7–51.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 99–120

The Modigliani-Miller Propositions After Thirty Years

Merton H. Miller

This issue of the *Journal of Economic Perspectives* appears on the 30th anniversary of the Modigliani-Miller propositions in "The Cost of Capital, Corporation Finance and the Theory of Investment," published in the *American Economic Review*, June 1958. The editors have invited me, if not to celebrate, at least to mark the event with a retrospective look at what we set out to do on that occasion and an appraisal of where the propositions stand today after three decades of intense scrutiny and often bitter controversy.

Some of these controversies can be now be regarded as settled. Our Proposition I, holding the value of a firm to be independent of its capital structure (that is, its debt/equity ratio) is accepted as an implication of equilibrium in perfect capital markets. The validity of our then-novel arbitrage proof of that proposition is also no longer disputed, and essentially similar arbitrage proofs are now common throughout finance.[1] Propositions analogous to, and often even called, M and M propositions, have spread beyond corporation finance to the fields of money and banking, fiscal policy and international finance.[2]

---

[1] Examples include Cornell and French (1983) on the pricing of stock index futures, Black and Scholes (1973) on the pricing of options and Ross (1976) on the structure of capital asset prices generally. For other, and in some respects, more general proofs of our capital structure proposition, see among others, Stiglitz (1974) for a general equilibrium proof showing that individual wealth and consumption opportunities are unaffected by capital structures; Hirshleifer (1965) and (1966) for a state preference, complete-markets proof; Duffie and Shafer (1986) for extensions to some cases of incomplete markets and Merton (forthcoming) for a spanning proof.

[2] See, for example, Wallace (1981) on domestic open-market operations; Sargent and Smith (1986) on central bank foreign-exchange interventions; Chamley and Polemarchakis (1984) on government tax and borrowing policies; and Fama (1980, 1983) on money, banking and the quantity theory.

■ *Merton H. Miller is Robert R. McCormick Distinguished Service Professor, Graduate School of Business, University of Chicago, Chicago, Illinois.*

Clearly Proposition I, and its proof, have been accepted into economic theory. Less clear, however, is the empirical significance of the MM value-invariance Proposition I in its original sphere of corporation finance.

Skepticism about the practical force of our invariance proposition was understandable given the almost daily reports in the financial press, then as now, of spectacular increases in the values of firms after changes in capital structure. But the view that capital structure is literally irrelevant or that "nothing matters" in corporate finance, though still sometimes attributed to us (and tracing perhaps to the very provocative way we made our point), is far from what we ever actually said about the real world applications of our theoretical propositions. Looking back now, perhaps we should have put more emphasis on the other, upbeat side of the "nothing matters" coin: showing what *doesn't* matter can also show, by implication, what *does*.

This more constructive approach to our invariance proposition and its central assumption of perfect capital markets has now become the standard one in teaching corporate finance. We could not have taken that approach in 1958, however, because the analysis departed too greatly from the then-accepted way of thinking about capital structure choices. We first had to convince people (including ourselves!) that there could be *any* conditions, even in a "frictionless" world, where a firm would be indifferent between issuing securities as different in legal status, investor risk and apparent cost as debt and equity. Remember that interest rates on corporate debts were then in the 3 to 5 percent range, with equity earnings/price ratios—then the conventional measure of the "cost" of equity capital—running from 15 to 20 percent.

The paradox of indifference in the face of such huge spreads in the apparent cost of financing was resolved by our Proposition II, which showed that when Proposition I held, the cost of equity capital was a linear increasing function of the debt/equity ratio. Any gains from using more of what might seem to be cheaper debt capital would thus be offset by the correspondingly higher cost of the now riskier equity capital. Our propositions implied that the *weighted average* of these costs of capital to a firm would remain the same no matter what combination of financing sources the firm actually chose.

Though departing substantially from the then-conventional views about capital structure, our propositions were certainly not without links to what had gone before. Our distinction between the real value of the firm and its financial packaging raised many issues long familiar to economists in discussions of the "money illusion" and money neutrality. Even some of the particular "financial illusions" to which we were directing attention had themselves already been noted by others as we duly cited in our paper. These earlier statements, however, had, to our knowledge, given rise to no follow-up work. The only prior treatment similar in spirit to our own was by David Durand (1952), who, as it turned out, also became our first formal critic (Durand, 1959). He had proposed, though not proved, as one of what he saw as two polar approaches to valuing shares, that investors might ignore the firm's then-èxisting capital structure and first price the whole firm by capitalizing its operating earnings *before* interest and taxes. The value of the shares would then be found by subtracting out the value of the bonds. But he rejected this possibility in favor of his other extreme, which he believed closer to the ordinary, real world way of valuing corporate shares, in which investors capitalized the firm's net income after interest and taxes with only a loose, qualitative adjustment for the degree of leverage in the capital structure.

That we too did not dismiss the seemingly unrealistic approach of looking through the momentary capital structure to the underlying real flows may well trace to the macroeconomic perspective from which we had approached the problem of capital structure in the first instance. Our main concern, initially, was with the determinants of *aggregate* economic investment by the business sector. The resources for capital formation by firms came ultimately from the savings of the household sector, a connection that economists had long found convenient to illustrate with schematic national income and wealth T-accounts, including, of course, simplified sectoral balance sheets such as:

Business Firms		Households	
Assets	Liabilities	Assets	Liabilities
Productive Capital	Debts owed to households	Debts of firms	Household net worth
	Equity in firms owned by households	Equity in firms	

Consolidating the accounts of the two sectors leads to the familiar national balance sheet:

Assets	Liabilities
Productive Capital	Household Net Worth

in which the debt and equity securities no longer appear. The value of the business sector to its ultimate owners in the household sector is thus seen clearly to lie in the value of the underlying capital. And by the same token, the debt and equity securities owned by households can be seen not as final, but only as intermediate assets serving to partition the earnings (and their attendant risks) among the many separate individual households within the sector.

Our value-invariance Proposition I was in a sense only the application of this macroeconomic intuition to the microeconomics of corporate finance; and the arbitrage proof we gave for our Proposition I was just the counterpart at the individual investor level of the consolidation of accounts, and the washing out of the debt/equity ratios at the sectoral level. In fact, one blade of our arbitrage proof had the arbitrager doing exactly that washing out. If levered firms were undervalued relative to unlevered firms, our arbitrager was called on to "undo the leverage" by buying an appropriate portion of both the levered firm's debt and its shares. On a consolidated basis, the interest paid by the firm cancels against the interest received and the arbitrager thus owned a pure equity stream. Unlevered corporate equity streams could in turn be relevered by borrowing on individual account if unlevered streams ever sold at a discount relative to levered corporate equity. That possibility of "homemade leverage" by individual investors provided the second and completing blade of our arbitrage proof of value invariance.

Our arbitrage proof drew little flak from those who saw it essentially as a metaphor—an expository device for highlighting hidden implications of the "law of one price" in perfect capital markets. But whether the operations we called arbitrage could *in fact* substitute for consolidation when dealing with real world corporations was disputed. Could investors, acting on their own, really replicate and, where required, wash out corporate capital structures? Even if they could not do so completely and immediately as in the formal proof, could they act completely enough and quickly enough to make the invariance proposition useful as a description of the central tendency in the real world capital market? These long-standing and still not completely resolved issues of empirical relevance of the MM propositions will be the primary focus in what follows here.

Three separate reasons (over and above the standard complaint that we attributed too much rationality to the stock market) were quickly offered by our critics for believing that individual investors could not enforce the corporate valuations implied by Propositions I and II. These lines of objection—relating to dividends, debt defaults, and taxes—each emphasized a different, distinctive feature of the corporate form of business organization; and each, in turn, will be reexamined here, taking full advantage of the hindsight of 30 years of subsequent research and events.

Before turning to these objections, however, it may be worth emphasizing at this point, while the sectoral tableaus are still near at hand to serve as a reminder, that our proposition that value was independent of capital structure at the individual firm level was never intended to suggest that the debt/equity ratio was *indeterminate*. At the firm level, there were clearly other costs of the various financial alternatives to be taken into account and we will get to some of these in due course. But even waiving the transaction and other second-order costs of financial packaging, the *aggregate* debt/equity ratio, unlike that of any individual firm where arbitrage governed, *would* reflect the risk preferences of households for holding wealth, including, needless to say, those preferences arising from life-cycle considerations. A general equilibrium macroeconomic model of the determination of aggregate real investment and the aggregate debt/equity ratio did appear as an appendix to the original working paper version of the article. But we found problems with one of the equations and removed the appendix from the version submitted for publication, intending to "get back to it someday." That day, alas, has yet to come, but some of the flavor at least of that general equilibrium treatment reappears in several later articles, notably Miller (1977), Miller and Scholes (1978), and Modigliani (1982).

Arbitrage, Dividends and the Corporate Veil

The law of one price is easily visualized in commodity settings where market institutions deliberately provide the necessary standardization and interchangeability of units. But to which of the many features of an entity as complex as an operating business firm would our financial equilibration extend?

We opted for a Fisherian rather than the standard Marshallian representation of the firm. Irving Fisher's view of the firm—now the standard one in finance, but then just becoming known—impounds the details of technology, production, and sales in a black box and focuses on the underlying net cash flow. The firm for Fisher was just an abstract engine transforming current consumable resources, obtained by issuing securities, into future consumable resources payable to the owners of the securities. Even so, what did it mean to speak of firms or cash flow streams being different, but still "similar" enough to allow for arbitrage or anything close to it?

Some of the answers would be provided, we hoped, by our concept of a "risk class," which was offered with several objectives in mind. At the level of the theory, it defined what today would be called a "spanning" set; the uncertain, underlying future cash flow streams of the individual firms within each class could be assumed perfectly correlated, and hence perfect substitutes. But the characteristics of those correlated streams could be allowed to differ from class to class. Hence, at the more practical level, the risk class could be identified with Marshallian industries—groupings around which so much academic and Wall Street research had always been organized. (Remember, in this connection, that the capital asset pricing models of Sharpe (1964) and Lintner (1965) and their later extensions that now dominate empirical research in finance had yet to come on the scene. For some glimpses of how more recent asset pricing frameworks can accommodate the MM propositions without reference to MM risk or MM arbitrage, see the comment by Stephen Ross that follows this article.) We hoped that the earnings of firms in some large industries such as oil or electricity generation might vary together closely enough not just for real world arbitragers to carry on their work of equilibration efficiently, but also to offer us as outside observers a chance of judging how well they were succeeding. Indeed, we devoted more than a third of the original paper (plus a couple of follow-up studies, notably Miller and Modigliani, 1966) to empirical estimates of how closely real world market values approached those predicted by our model. Our hopes of settling the empirical issues by that route, however, have largely been disappointed. Direct statistical calibration of the goodness of fit of the MM value-invariance propositions has not so far been achieved by us or others for a variety of reasons, some of which will be noted further in due course below.

Investor Arbitrage When Dividends Differ: The Dividend Invariance Proposition

Although the risk class, with its perfect correlation of the underlying real cash streams, may have provided a basis for the arbitrage in our formal proof, there remained the sticking point of how real world market equilibrators could gain access to a firm's operating cash flows, let alone to two or more correlated ones. As a matter of law, what the individual equity investor actually gets on buying a share is not a right to the firm's underlying cash flow, but only to such cash dividends as the corporation's directors choose to declare. Must these man-made payouts policies also be assumed perfectly correlated along with the underlying cash flows to make the equilibration effective? If so, the likely empirical range of the value-invariance proposition would seem to be narrow indeed.

A second MM invariance proposition—that the value of the firm was independent of its dividend policy—was developed in part precisely to meet this class of objections. The essential content of the dividend irrelevance argument was already in hand at the time of the original leverage paper and led us there to dismiss the whole dividend question as a "mere detail"—not the last time, alas, that we may have overworked that innocent word "mere." We stated the dividend invariance proposition explicitly, and noted its relation to the leverage proof in the very first round of replies to our critics (Modigliani and Miller, 1959, esp. pp. 662–668). But because dividend decisions were controversial in their own right and because considering them raised so many side issues of valuation theory and of practical policy, both private and public, we put off the fuller treatment of dividends to a separate paper that this time really did appear, but only in 1961, some three years after the first one.

That the close connection in origin of the two invariance propositions has not been more widely appreciated traces not only to their separation in time, but probably also to our making no reference to arbitrage (or even to debt or equity) in the proof of the dividend invariance proposition. Why bring in arbitrage, we felt, when an even simpler line of proof would serve? The dividend invariance proposition stated only that *given* the firm's investment decision, its dividend decision would have no effect on the value of the shares. The added cash to fund the higher dividend payout must come from somewhere, after all; and with investment given, that somewhere could only be from selling off part of the firm. As long as the securities sold off could be presumed sold at their market-determined values, then, whether the analysis was carried out under conditions of certainty or uncertainty, the whole operation of paying dividends, investment given, could be seen as just a wash—a swap of equal values not much different in principle from withdrawing money from a passbook savings account.

The informational content of dividends. Managerial decisions on dividends thus might affect the cash component of an investor's return; but they would not affect the *total* return of cash plus appreciation, and the total is what mattered. In practice, of course, even changing the cash-dividend component often seemed to matter a great deal, at least, to judge by the conspicuous price jumps typically accompanying announcements of major boosts or cuts in dividends. These highly visible price reactions to dividend announcements were among the first (and are still the most frequently mentioned) of the supposed empirical refutations of the MM value-invariance principle. By invoking the dividend invariance proposition to support the leverage invariance proposition, we seemed to have succeeded only in substituting one set of objections for another.

But, as we suggested in the 1961 dividend paper, these price reactions to dividend announcements were not really refutations. They were better seen as failures of one of the key assumptions of both the leverage and dividend models, namely, that all capital market participants, inside managers and outside investors alike, have the same information about the firm's cash flows. Over long enough time horizons, that all-cards-on-the-table assumption might, we noted, be an entirely acceptable approximation, particularly in a market subject to Securities and Exchange Commission disclosure rules. But new information is always coming in; and over shorter runs, the firm's inside managers were likely to have information about the firm's prospects not yet known to or fully appreciated by the investing public at large. Management-initiated actions on dividends or other financial transactions might then serve, by implication, to convey to the outside market information not yet incorporated in the price of the firm's securities.

Although our concern in the 1961 dividend paper was with the observed announcement effects of dividend decisions, informational asymmetry also raised the possibility of strategic behavior on the part of the existing stockholders and/or their management agents. Might not much of the price response to dividend (and/or other capital-structure) announcements simply be attempts of the insiders to mislead the outsiders; and if so, what point was there to our notion of a capital market equilibrium rooted solely in the fundamentals? Our instincts as economists led us to discount the possibility that firms could hope to fool the investing public systematically; but at the time, we could offer little more support than a declaration of faith in Lincoln's Law—that you can't fool all of the people all of the time. By the 1970s, however, the concept of an information equilibrium had entered economics, and it soon came to the field of corporate finance as well. Bhattacharya (1979) noted the formal similarity between Spence's (1973) job-market signalling model and the MM dividend model with asymmetric information. Ross (1977) showed how debt/equity ratios might also serve to signal, in the technical sense, managements' special information about the firm's future prospects. The extent to which these and subsequent asymmetric information models can account for observed departures from MM invariance has not so far been convincingly established, however.[3]

The interaction of investment policy and dividend policy. The dividend invariance proposition, as we initially stated it, highlights still another way in which the corporate form of organization, and especially the separation it permits between ownership and management, can have effects that at first sight at least seem to contradict the MM value-invariance predictions. Recall that the dividend invariance proposition takes the firm's investment decision as given, which is just a strong way of saying that the level of investment, whatever it might be, is set by management *independently* of the dividend. Without imposing such an "other-things-equal" condition, there would, of course, be no way of separating the market's reaction to real investment events from reaction to the dividend and any associated purely financial events.

---

[3] For a survey of recent results on dividend signalling see Miller (1987). For a survey of asymmetric information models in finance more generally see Stiglitz (1982).

In the real world, of course, the financial press reports single-company stories, not cross-sectional partial regression coefficients. In these single-company tales, the investment decision and the dividend/financing decisions are typically thoroughly intertwined. But if the tale is actually one of a firm cutting back unprofitable investments and paying out the proceeds as dividends, followed by a big run-up in the firm's price, then the MM invariance proposition may seem to be failing, but it is really not being put to the test. Nor is this scenario only hypothetical. Something very much like it appears in a number of the most notorious of recent takeover battles, particularly in the oil industry where some target firms had conspicuously failed to cut back their long-standing policies of investment in exploration despite the drastic fall in petroleum prices.

In a sense, as noted earlier, these gains to shareholders from ending a management-caused undervaluation of the firm's true earning power can also be viewed as a form of capital market arbitrage, but not one that atomistic MM investors or arbitragers can supply on their own. Once again, the special properties of the corporate form intrude; this time the issue is the voting rights that attach to corporate shares and the majority rules (and sometimes supermajority rules) in the corporate charters that determine the control over the firm's decisions. Much of the early skepticism, still not entirely dispelled, about the real empirical force of interfirm arbitrage, MM-arbitrage included, traces to these properties of corporate shares beyond their purely cash-flow consequences. A particular example of the obstacle they offered to effective capital market equilibration was that of closed-end investment funds. In 1958, as still today, closed-end funds often sold in the market at a substantial discount to net asset value—a discount the could be recaptured by the shareholders merely (that word, again!) by getting enough of them to vote to convert to open-end fund status.

Voting rights were one, but only one, of the special properties of corporate common shares from which the MM model abstracted. There was another, namely limited liability, which appeared to loom even larger as an obstacle to the "homemade" leverage on which the arbitrage proof relied. That was the second of the early lines of objection to the invariance proposition and its proof. As is turns out, limited liability is not really critical either to the proof or to the empirical force of the invariance proposition. But it may have been made to seem so at the time by an unfortunate tactical simplification we made in presenting the original proof.

MM Invariance with Limited Liability and Risky Debt

The troublesome tactical simplification in the original proof was our taking bonds or other debt instruments to be securities not merely of lower risk than common stocks, but of no risk whatever. The promises their issuers made would be honored. Drawing so sharp a line between risky stocks and riskless bonds served, we thought, to bring out the risks of corporate leveraging as such, and, to that extent, also to explain how the seeming gains from using cheap debt can be offset by the higher risks and hence costs of leveraged equity (our MM Proposition II) keeping the weighted average risks and costs the same (our Proposition I). But making bonds riskless also made all debts effectively indistinguishable, thereby, leaving corporate finance, in the strict sense, with nothing to do. Individual investors, creating or eliminating leverage with debts on their personal accounts, could always match or undo any capital structure decision at the corporate level. Thus, ironically, the riskless debt assumption we introduced orirginally to sharpen the line between corporate stocks and bonds seemed to have blurred the line between the corporation and other forms of business organization.

Arbitrage When Corporations, But Not Individual Investors, Have Limited Liability

Limiting the liability for the firm's debts to the initial capital contribution was, after all, one of the presumed advantages that had led the owner-organizers of capital-intensive enterprises to substitute the corporate form for personal proprietorships or partnerships in which their own personal wealth would have been open to unsatisfied creditors. Hence individuals could not really hope to duplicate corporate structures on their own as the original MM arbitrage proof seemed to require, unless —to avoid any chance of default—further and unrealistically severe restrictions were imposed both on the firm's worst-outcome cash flow prospects and on the amounts of debt it could issue.

In defense of our simplification let it be noted at least that the notion of an idealized riskless rate of interest had long been, and still remains, in the standard tool-kit of economic theory. We also had no qualms about gaps that our riskless debt restriction might open in the spectrum of risks available to investors. Even if our firms did not fill the zone between riskless debts and risky equities with low, but more-than-riskless-instruments, investors could always span it by mixing stocks with riskless bonds in their personal portfolios.

There was, moreover, one sense, though a limited one, in which both the value-invariance proposition and the arbitrage proof would go through even though shareholders in levered corporations, but not individuals levered on private account, had limited liability in the event of a default. That institutional difference in liability need not be critical so long as the potential arbitragers in the household sector were substantial net creditors to the corporate sector. They could then bring about the equivalent of any required increase in limited-liability leveraging of unlevered shares merely by selling off some of the risky corporate bonds held in their personal portfolios. The same sort of argument could be invoked, as we noted, not just for risky corporate bonds but for other securities that are specific to corporate issuers, such as preferred stocks.

Corporations as Arbitragers

An even more effective way of finessing the limited-liability issue and, in the process, of emphasizing that the assumed risklessness of the debt was *not* the key to the invariance result or its arbitrage proof, would have been to permit corporations generally, and not just individuals, to participate in the equilibrating arbitrage operations. Such an extension of the set of eligible arbitragers (suggested, in fact, by some early commentators on our paper, notably Stiglitz (1969, 1974)), may perhaps have escaped us at the time because of the mind set induced by our sectoral consolidation tableaus in which firms and households were assigned such fundamentally distinctive economic roles. Had we drawn those roles less sharply and let corporations hold securities as assets as well as issue them as liabilities (as in the models of Duffie and Shafer, 1986), we might have spared the profession much needless subsequent controversy. The real world, needless to say, has always been far less fastidious than we were in sharply differentiating the functions of firms and households; much of what passes these days for corporate-raiding-cum-restructuring is just MM leverage arbitrage, but channeled through the raider's corporate, rather than personal, investment account.

Corporations and Financial Innovation

With risky debt and active corporate players, the general equilibrium issue of invariance of real allocations to financial policies must also be reopened, because some prospect then arises for "creative" corporate finance. Through new, high leverage, limited-liability securities, firms might enhance their own value (and the social welfare) by offering risk and return combinations that fully-liable investors could not hope to achieve on their own.

Such new, high risk opportunities might well have been among the first sources to be looked to had the recent spectacular gains from corporate restructuring occurred in the late 1960s or early 1970s. High leverage was far from easy to come by in those days. And even today, the scarcity of long-term ultra-high leverage opportunities may perhaps account for some of the otherwise perplexingly high prices for the shares—or "stubs" as they are sometimes disparingly called—left over after some recent massive debt-financed corporate share repurchases. But it is hard to see any major role in the gains from recent restructurings by U.S. industrial corporations that can be traced to the creation of limited liability, high-risk investment opportunities by ordinary corporations in a form not otherwise attainable by individual investors. That some gains have arisen from "expanding the state space" with new kinds of financial instruments and financial repackagings may well be true, of course. An example of such market-completing repackagings might be Collateralized Mortgage Obligations, where the uncertain, interest-rate-driven repayments of the underlying household mortgages may have created a new kind of interest rate risks, and hence of interest rate hedges, not found in the normal run of debt instruments. (Some cynics might be more inclined to credit any gains from CMO's to opening holes in the regulatory net covering the savings and loan sector.) But with few exceptions (among them, perhaps, so-called auction-rate preferred stocks and some recent securities linked to commodities and exchange rates) the panoply of corporate securities is pretty much what it always has been. The main financial innovations of recent years have come not from operating corporations, but from the commodity and securities exchanges which have created an entirely new noncorporate family of limited-liability instruments for investors. With these instruments has also come a new way of looking at corporate capital structures and the MM propositions.

The Options Interpretation of Levered Stock

The new instruments offering individual investors low cost, limited-liability substitutes for corporate leverage were, of course, put and call options and their many variations and permutations. The mechanics of option theory are by now too well-known among economists to warrant extensive review here, but the insights the existence of options gives to the economic (as opposed to the legal) view of debt, and especially risky debt, are perhaps still worth a further word or two.[4]

The legal view of debt, reflected also in the everyday usage of the term, involves a long laundry list of attributes whose essence is obligation. Debts are *promises* by the debtor to pay specific amounts on specific dates. If the promise is not kept, the creditor can sue.

The restrictiveness of any definition of debt relying mainly on the notions of obligation and legal remedies becomes clear immediately on confronting the complexities of real world capital structures. Where, to name just one of a hundred forms of fancy-featured corporate securities, is something even as simple and familiar as preferred stock to be placed? For the common stockholders, the fixed-maximum dividend on preferred shares clearly provides leverage in exactly the same way as would a fixed-interest bond. Yet in law (including tax law) and in accounting, preferred stock is an *equity* security since it carries no enforceable obligation to pay. The discretion to pay preferred dividends is normally left entirely to the corporation's directors, as it is with dividends on common. The directors will choose to pay the preferred dividends only if the costs to common shareholders of not paying them—mainly delay (possibly long) of cash dividends on the common, possibly some reduction in voting power and certainly some loss of reputation capital for future financings—exceed those of honoring the promise. But that is true as much for securities that are legal debts as it is for preferred shares. The firm pays its debts not just because the law says it must, but because the value of the stock to its shareholders is greater to them if the firm pays the debts than if it doesn't. Otherwise the shareholders can default on the payment, invoke limited liability and turn the keys over to the bondholders.

The shareholders in a levered, limited-liability corporation are thus essentially the owners of a call option whose exercise price is the promise in the senior securities. The academic finance profession is still exploring the full ramifications of this important insight, which traces along with so much else to the original Black-Scholes option paper (1973). That it has already thrown new light on the MM propositions is hardly surprising since MM-type arbitrage arguments were explicitly invoked by Black and Scholes in deriving their option valuation formula. A key to their derivation was their recognition that a stock and an option on that stock were in the same "risk class" (that is, they were prefectly correlated) at least for local variations in their prices; and that they could be kept in the same risk class in the face of larger price variations by dynamic rebalancing of the number of options over time.

---

[4] An interpretive review of the basics of option theory is given by Rubenstein (1987) in an earlier issue of this journal.

The put-call parity theorem as an MM proposition. The closeness of the connection between modern option theory and the earlier MM analysis stands out even more strikingly when attention turns to the relations between the prices of different options on the same stock. The familiar Put-Call Parity Theorem (first stated formally in Stoll, 1969), is really nothing more than the MM Proposition I in only a mildly concealing disguise!

To see this, recall the put-call parity requires that, under frictionless conditions, the following relation must hold as an equality for any value of K:

$$S = C(K) + Ke^{-rt} - P(K)$$

where S is the current price of the stock against which the put and call options are to be written, K the common exercise price of the options, T the time to maturity of the options, r the riskless rate of interest, and $C(K)$ and $P(K)$ the current prices of the call and put options respectively. Now reinterpret each term in the light of the Black-Scholes capital structure analogy. (For simplicity, pretend that the firm has outstanding only a single, zero coupon bond. More general cases are considered in Merton (1974) and Black and Cox (1976).) S becomes the value of the firm's cash flow; K the face or book value of the firm's liabilities; and $C(K)$ the market value of the levered shares in the firm. Ke^{-rT} is the market value that the firm's debt would have if it were riskless. But of course, it is not riskless. Should the value of S at maturity of the debt turn out to be less than K, the shareholders will invoke limited liability and put the firm back to the creditors. The actual market value of the debt is thus only $Ke^{-rT} - P(K)$, the present face value of the debt *less* the value of the shareholders' put. The market value both of the firm's debts and its equity are thus clearly functions of the firm's leverage. But by the Put-Call Parity Theorem as shown in the equation, their *sum* is independent of leverage. The combined pieces of the capital structure always add up to S, the value of the underlying unlevered cash flow exactly as in the MM Proposition I.

Not the least of the gains from invoking the Put-Call Parity analogy is the light it throws on the problems of establishing empirically the goodness of fit of the MM Propositions. The fit of Put-Call Parity can be calibrated because all its terms are directly observable. But the "true" value of a firm—the MM concept that corresponds to the S in the equation describing of the Put-Call Parity Theorem—or the "risk class" to which the firm belongs, is not observable.

Even without formal statistical testing, however, we have no shortage of potential candidates for forces that might well lead the market to depart systematically and persistently from the predictions of the original MM value-invariance propositions. One such likely candidate, the third of the original lines of objection, has loomed so large in fact as to have dominated academic discussions of the MM propositions, at least until the recent wave of corporate takeovers and restructurings became the new focus of attention. That candidate is the corporate income tax, the one respect in which everyone agreed that the corporate form really did matter.

The MM Propositions in a World with Taxes

The U.S. Internal Revenue Code has long been the classic, and by now is virtually the world's only, completely unintegrated tax system imposing "double taxation" of corporate net income. A separate income tax is first levied directly in the corporation and, except for certain closely-held corporations who may elect to be taxed like partnerships under Subchapter S of the Code, a second tax is then levied at the personal level on any income flows such as dividends or interest generated at the corporate level. Double taxation of the interest payments is avoided because interest on indebtedness is considered a cost of doing business and hence may be deducted from corporate gross income in computing net taxable corporate earnings. But no such allowance has been made for any costs of equity capital.[5]

If the separate corporate income tax were merely a modest franchise tax for the privilege of doing business in corporate form, as was essentially the case when it was introduced in the early years of this century, the extra burden on equity capital might be treated as just one more on the long list of second-order differences in the costs of alternative sources of capital for the firm. But at the time of our 1958 article the marginal tax rate under the corporate income tax had been close to and sometimes over 50 percent for nearly 20 years and it remained there for almost another 30 years until dropped to 34 percent by the Tax Reform Act of 1986. Cost differentials of this size were just too big to be set aside in any normative or empirical treatments of real world capital structure choices.

Strictly speaking, of course, there is one sense, albeit a somewhat strained one, in which the basic value-invariance does go through even with corporate taxes. The Internal Revenue Service can be considered as just another security holder, whose claim is essentially an equity one in the normal course of events (but which can also take on some of the characteristics of secured debt when things go badly and back taxes are owed). Securities, after all, are just ways of partitioning the firm's earnings; the MM propositions assert only that the sum of the values of all the claims is independent of the number and the shapes of the separate partitions.

However satisfying this government-as-a-shareholder view may be as a generalization of the original model, the fact remains that the government, though it sometimes gives negative taxes or subsidies for some kinds of investment, does not normally buy its share with an initial input of funds that can serve to compensate the other stockholders for the claims on income they transfer to the Treasury. Nor are we talking here of taxation-according-to-the-benefits or of the rights of eminent domain, or even of whether the corporate tax might ultimately be better for the shareholders or for the general public than alternative ways of raising the same revenue. For the nongovernment equity claimholders, the government's claim to the firm's earnings is a net subtraction from their own.

---

[5] Two exceptions should be noted for the record. An undistributed profits tax from which dividends were deductible was in force for two years in the late 1930s. The excess profits tax during World War II also allowed a deduction not for dividends, but for the "normal profits" of the firm.

The MM Tax-adjusted Leverage Proposition

Allowing for that subtraction can lead to a very different kind of MM Proposition, though one, as we showed in our Tax Correction article (1963) that can still be derived from an arbitrage proof along lines very similar to the original. This time, however, the value of the firm (in the sense of the sum of the values of the private, nongovernmental claims) is *not* independent of the debt/equity division in the capital structure. In general, thanks to the deductibility of interest, the purely private claims will increase in value as the debt ratio increases. In fact, under conditions which can by no means be dismissed out of hand as implausible, we showed that the value of the private claims might well have no well-defined interior maximum. The optimal capital structure might be all debt!

In many ways this tax-adjusted MM proposition provoked even more controversy than the original invariance one, which could be, and often was, shrugged off as merely another inconsequential paradox from some economists' frictionless dreamworld. But this one carried direct and not very flattering implications for the top managements of companies with low levels of debt. It suggested that the high bond ratings of such companies, in which the management took so much pride, may actually have been a sign of their incompetence; that the managers were leaving too much of their stockholders' money on the table in the form of unnecessary corporate income tax payments, payments which in the aggregate over the sector of large, publicly held corporations clearly came to many billions of dollars.

We must admit that we too were somewhat taken aback when we first saw this conclusion emerging from our analysis. The earlier modeling of the tax effect in our 1958 paper, which the 1963 paper corrected, had also suggested tax advantages in debt financing, but of a smaller and more credible size. By 1963, however, with corporate debt ratios in the late 1950s not much higher than in the low tax 1920s (see Miller, 1963) we seemed to face an unhappy dilemma: either corporate managers did not know (or perhaps care) that they were paying too much in taxes; or something major was being left out of the model. Either they were wrong or we were.

The offsetting costs of debt finance. Much of the research effort in finance over the next 25 years has been spent, in effect, in settling which it was. Since economists, ourselves included, were somewhat leerier then than some might be now in offering mass ineptitude by U.S. corporate management as an explanation for any important and long-persisting anomalies, attention was naturally directed first to the possibly offsetting costs of leveraging out from under the corporate income tax. Clearly, leveraging increased the riskiness of the shares, as we ourselves had stressed in our original Proposition II and its tax-adjusted counterpart. A sequence of bad years, moreover, might wipe out the firm's taxable income, and, given the very ungenerous treatment of losses in our tax law, that could reduce, possibly quite substantially, any benefits from the interest tax shields. A run of very bad years might actually find a highly-levered firm unable (or, as the option theorists might prefer, unwilling) to meet its debt service requirements, precipitating thereby any of the several processes of recontracting that go under the general name of bankruptcy. These renegotiations can be costly indeed to the debtor's estate, particularly when many separate classes of creditors are involved.[6]

The terminal events of bankruptcy are not the only hazards in a high debt strategy. Because the interests of the creditors and the stockholders in the way the assets are managed need not always be congruent, the creditors may seek the additional protection of restrictive covenants in their loan agreement. These covenants may not only be costly to monitor but may foreclose, if only by the time delay in renegotiating the original terms, the implementation of valuable initiatives that might have been seized by a firm less constrained. Nor should the transaction and flotation costs of outside equity financing be neglected, particularly in the face of information asymmetries. Prudence alone might thus have seemed to dictate the maintenance of a substantial untapped reserve of quick borrowing power, especially in an era when those managing U.S. corporations (and the financial institutions buying their debt securities) still had personal memories of the debt refinancing problems in the 1930s.

We dutifully acknowledged these well-known costs of debt finance, but we were hard put at the time to see how they could overweigh the tax savings of up to 50 cents per dollar of debt that our model implied. Not only did there seem to be potentially large amounts of corporate taxes to be saved by converting equity capital to deductible-interest debt capital, but there appeared to be ways of doing so that avoided, or at least drastically reduced, the secondary costs of high-debt capital structures. The bankruptcy risk exposure of junior debt could have been blunted with existing hybrid securities such as income bonds, to take just one example, under which deductible interest payments could be made in the good years, but passed or deferred in the bad years without precipitating a technical default. For reducing the moral hazards and agency costs in the bondholder-stockholder relation, the undoing-of-leverage blade in the original MM proof offered a clue: let the capital suppliers hold some of each, either directly or through convertible or exchangeable securities of any of a number of kinds. In sum, many finance specialists, myself included, remained unconvinced that the high-leverage route to corporate tax savings was either technically unfeasible or prohibitively expensive in terms of expected bankruptcy or agency costs.

Junk Bonds, Leveraged Buyouts, and the Feasibility of High Leverage Strategies

A number of recent developments in finance can be seen as confirming the suspicions of many of us academics in the early 1960s that high leverage strategies to reduce taxes were indeed entirely feasible. Among these, of course, is the now large outstanding volume of what are popularly known as "junk bonds." The very term is a relic of an earlier era in which the distinguishing characteristic of bonds as investments was supposedly their presence at the low-risk end of the spectrum. High-risk, high-yield bonds did exist, of course, but were typically bonds issued initially with high ratings by companies that had subsequently fallen on hard times. The significant innovation in recent years—and it is still a puzzle as to why it took so long—has been in the showing that, contrary to the conventional wisdom, junk bonds could in fact be issued and marketed successfully by design, and not just as "fallen angels."

---

[6] The perceived complexity of the present bankruptcy code (and perhaps even the very reason for having such a code) reflect mainly the need for resolving conflicts within and between the various classes of creditors. The difficulties parallel those encountered elsewhere in "common pool" problems (Jackson, 1986).

The designs utilizing new risky-debt securities have often taken the very conspicuous form of "leveraged buyouts" of the outside shareholders by a control group typically led by the existing top management. The device itself is an old one, but had been confined mainly to small firms seeking both to assure their continuity after the death or retirement of the dominant owner-founder, and to provide more liquidity for the entrepreneur's estate. The new development of recent years has been the ability, thanks in part to the market for junk bonds, to apply the technique to a much wider range of publicly held big businesses with capitalizations now routinely in the billions, and with new size records being set almost every year.

The debt/equity ratios in some recent LBO's have reached as high as 9 to 1 or 10 to 1 or even more, far beyond anything we had over dared use in our numerical illustrations of how leverage could be used to reduce taxes. The debtor/creditor incentive and agency problems that might be expected under such high leverage ratios have been kept manageable partly by immediate asset sales, but over the longer term by "strip financing"—trendy investment-banker argot for the old device of giving the control and most of the ownership of the equity (except for the management incentive shares) to those providing the risky debt (or to the investment bankers they have designated as monitors). The same hold-both-securities approach, as in our arbitrage proof, has long been the standard one in Japan where corporate debt ratios are, or are at least widely believed to be, substantially higher than for their U.S. counterparts.

Some possible non-tax gains from leveraging. The recent surge of leveraged buyouts not only shows the feasibility of high leverage capital structures for reducing corporate income taxes, but also suggests at least two other possible sources for the gains to the shareholder that may accompany a major recapitalization with newly-issued debt. The firm may, for example, already have had some long-term debt outstanding when the additional debt needed to accomplish the buyout was arranged. Even in a no-tax world the "no gain from leverage" implication of the original MM invariance proposition might fail if the new debt was not made junior in status to the old, if the old bond covenant was "open ended," as many still are, and if the new bonds were issued under it. Assuming no change in the underlying earning power from the recapitalization, the original creditors would then find the value of their claim diluted. The benefits of this dilution of the old bondholders accrue, of course, to the stockholders, which is why it has often been labeled "theft," particularly by the adversely affected bondholders. Finance specialists prefer the less emotionally charged term "uncompensated wealth transfer."

The high debt ratios in LBO's also redirect attention to the assumption, shown earlier to be crucial to the MM dividend invariance proposition, that the firm's financial decisions can be taken as independent of its real operating and investment decisions. That assumption never sits well and certainly the notion that heavy debt burdens might indeed lead to overcautious business behavior has long been part of the folk wisdom on the dangers of debt. The new wrinkle to the interdependence argument brought in recently by the defenders of LBO's has been to stress the positive *virtues* of having managers face large debt obligations. Managements in such firms must work hard and diligently indeed to achieve any earnings above interest to enhance the value of the residual equity they hold in the firm. By accepting such heavy debt-service burdens, moreover, the managers are making a binding commitment to themselves and to the other residual equity holders against yielding to the temptations, noted earlier, to pour the firm's good money down investment rat holes.[7]

Voluntary recapitalizations and the MM *dividend proposition.* High debt ratios have been installed in some U.S. firms in recent years not just by outside-initiated LBO's but through voluntary recapitalizations. Sometimes, it is true, the motivation has been merely to fend off an imminent hostile takeover, but sometimes the tax benefits have been very clearly emphasized. Even apart from the tax angles, nothing in the practice of finance these days could be more quintessentially MM than these often highly visible "self takeovers," as some wag has dubbed them. Leverage-increasing recapitalizations of this kind do indeed raise the firm's debt/equity ratio, but because the proceeds of the new bonds floated are turned over to the shareholders, the self takeovers also reunite in a single operation the two Siamese-twin MM propositions, the leverage proposition and the dividend proposition, joined together originally at birth, but soon parted and living separate lives thereafter.

The dividend proposition, as noted earlier, was put forward initially to overcome a line of objection to the leverage proof. But how dividends might actually affect real world prices raises other issues, which in turn have led to as much controversy and to an even larger number of discordant empirical findings. Once again, moreover, major tax differentials intruded, this time the gap between rates on dividends and capital gains under the personal income tax, again with what seemed in the late 1950s and early 1960s to have strikingly unorthodox policy implications. Some high income stockholders clearly would have been better off if the firm paid no dividends and simply reinvested its earnings or bought shares in other corporations. That much every real-world conglomerator and every public finance specialist surely knew. But the value-for-value presumption of the MM dividend proposition carried within it some further advice. There were better ways to avoid taxes on dividends than pouring the firm's money down rat holes; for one, use the money to buy back the firm's shares! For the taxable shareholders, buybacks at market-determined prices could transform heavily-taxed dividends into less-heavily taxed capital gains, and, better yet, into unrealized capital gains for shareholders who choose not to sell or trade their shares. Unlike a declared regular dividend, moreover, an announced share repurchase, whether by tender or by open market purchases, carried no implied commitments about future payouts.

---

[7] This view of debt service as a device for reining in managerial discretion is a major strand in what has come to be called the "free cash-flow" theory of corporate finance. For an account of that theory in an earlier issue of this journal, see Jensen (1988).

Personal-corporate Tax Interactions and Capital Market Equilibrium

These tax-advantaged dividend-substitution properties of share repurchase may also offer a clue as to why the leveraging of corporate America out from under the corporate income tax may have been so long delayed. The point is not so much that share repurchase by itself has been a major vehicle deliberately invoked by corporations to reduce the personal income taxes of their shareholders, though its potential for that purpose certainly has not been lost on corporate treasurers and directors.[8] But the very presence of such a possibility at the corporate level serves as a reminder that the U.S. tax system has not one but two distinct taxes that bear on capital structure choices. Any model of capital market equilibrium must allow for both, and for their interactions.

In particular, under reasonable assumptions, the joint corporate-personal tax gains from corporate leverage, G_L can be expressed in the following relatively transparent formula (see Miller, 1977):

$$G_L = \left[1 - \frac{(1 - t_c)(1 - t_{PS})}{1 - t_{PB}}\right] B_L$$

where B_L is the value of the levered firm's interest-deductible debts, t_c is the marginal corporate tax rate and t_{PS} and t_{PB} are the marginal investor's personal marginal tax rates respectively on income from corporate shares and income from interest-bearing corporate debts. In the special case in which the personal income tax makes no distinction between income from debt or from equity (that is, $t_{PS} = t_{PB}$), the gain from leverage reduces to $t_c B_L$, precisely the expression in the MM tax model.[9] But in the contrasting extreme special case in which (a) the capital gains provisions or other special reliefs have effectively eliminated the personal tax on equity income, (b) full loss offsets are available at the corporate level and (c) the marginal personal tax rate on interest income just equals the marginal corporate rate, the purely tax gains from corporate leverage would vanish entirely. The gains from interest deductibility at the corporate level would be exactly offset by the added burden of interest includability under the personal tax—an added burden that in equilibrium would be approximated by risk-adjusted interest rate premiums on corporate and Treasury bonds over those on tax-exempt municipal securities.

---

[8] Most economists, upon first hearing about share repurchase as an alternative to dividend payments, assume that the Internal Revenue Service must surely have some kind of magic bullet for deterring so obvious a method of tax avoidance. (See Bhattacharya, this issue.) It doesn't; or at least not one that will work in the presence of even minimally competent tax lawyers.

[9] That special case assumes, among other things, that debt, once in place, is maintained or rolled over indefinitely. For valuing the tax savings when debts are not perpetuities, see the comment on this paper by Franco Modigliani in this issue.

This somewhat surprising special case of zero net gain from corporate leverage has inevitably received the most attention, but it remains, of course, only one of the many potentially interesting configurations for market equilibrium. Stable intermediate cases are entirely possible in which some gains to corporate leverage still remain, but thanks to the capital gains or other special provisions driving t_{PS} below t_{PB}, or to limitations on loss offsets, those gains at the corporate level are substantially below those in the original MM tax model. The tax gains from leverage might, in fact, even be small enough, when joined with reasonable presumed costs of leverage, to resolve the seeming MM anomaly of gross underleveraging by U.S. corporations. For some recent empirical tests of such an intermediate equilibrium using the premium over municipals, see Buser and Hess (1986). Kim (1987) offers a wide-ranging survey of recent theoretical and empirical research on capital market equilibrium in the presence of corporate-personal income tax interactions.

The MM Propositions and the Recent Tax Reform Act

Any such "Debt and Taxes" equilibrium, however, that the corporate sector might have reached in the early 1980s by balancing costs of debt finance against MM tax gains from leverage must surely have been shattered by the Tax Reform Act of 1986. That act sought (among other things) to reverse the long steady slide, accelerating in the early 1980s, in the contribution of corporate income taxes to total federal tax revenues. But in attempting to increase the load on corporations, Congress seemed to have overlooked some of the interactions between corporations and individual investors that lie at the heart of the MM propositions and their later derivatives. For shareholders taxable at high marginal rates on interest or dividends under the personal income tax, for example, maintaining assets in corporate mode and suffering the corporate tax hit might make sense, provided enough of the after-corporate-tax earnings could be transmuted to long-deferred, low-taxed capital gains by profitable reinvestment in real assets. In fact, over much of the life of the income tax, when shares were held largely by wealthy individuals and hardly at all by pension funds or other tax-exempt holders, the corporate form of organization for business with great growth potential may well have been the single, most important tax shelter of all.

But the pattern of tax advantages that encouraged the encouraged the accumulation of wealth in corporate form appears to have been altered fundamentally by the Tax Reform Act of 1986. The investment tax credit and related tax subsidies to fixed investment have been phased out. The marginal rate on the highest incomes under the personal income tax has now been driven to 28 percent and hence below the top corporate rate of 34 percent. The long-standing personal income tax differential in favor of long-term realized capital gains has been eliminated, though income in that form still benefits from a variety of timing options and from the tax-free write-up of any accumulated gains when the property passes to heirs. The analogous tax-free write-up privileges for corporate deaths or liquidations, however, formerly allowed under the so-called *General Utilities* doctrine, have now been cut back by the TRA and some of its recent predecessors, reducing still further the tax benefits of the corporate form.

To finance specialists familiar with the MM propositions, these combined changes suggest that Congressional hopes of substantially increasing the yield of the corporate income-tax—that is to say, their hopes of reinstating the double taxation of corporate profits—may well be disappointed. (For some recent signs of congressional concerns on this score see Brooks (1987) and Canellos (1987).) Our capital markets and legal institutions offer too many ways for averting the double hit. Corporations can split off their cash-cow properties into non-corporate "flow-through" entities. And, as has been the running theme of this entire section, firms retaining corporate form can always gut the corporate tax with high leverage capital structures. In fact, under not entirely implausible conditions (notably that the marginal bondholder is actually a tax-exempt pension fund rather than a taxable individual investor, implying that the t_{PB} in the gains from leverage equation is zero), the incentive to leverage out the corporate tax may now actually be as high or higher than it was back in 1963. The statutory top corporate tax rate has indeed been cut; but with the investment tax credit and accelerated depreciation also blown away by the Tax Reform Act of 1986, many capital-intensive corporations may now, for the first time in a very long while, be facing the unpleasant prospect of actually paying substantial corporate taxes.

Perhaps that observation can serve as a fitting note of uncertainty, or at least of unfinished business on which to close this look back at the MM propositions. The open questions about those propositions have long been the empirical ones, as noted here at many points. Are the equilibria implied by the propositions really strong enough attractors to demand the attention of those active in the capital markets either as practioners or as outside observers? In the physical or biological sciences one can often hope to answer such questions by deliberately shocking the system and studying its response. In economics, of course, direct intervention of that kind is rarely possible but nature, or at least Congress, can sometimes provide a substitute. The U.S. tax system is a pervasive force on business decisions of many kinds, but especially so on the class of financial decisions treated in the MM propositions. Tax considerations have for that reason always figured prominently in the field of finance. Occasionally, the profession may even see changes in the tax regime drastic enough for the path of return to a new equilibrium to stand out sharply against the background of market noise. Whether the Tax Reform Act of 1986 is indeed one of those rare supershocks that can validate a theory remains to be seen.

■ *Helpful comments on an earlier draft were made by George Constantinides, Melvin Reder, Lester Telser, Hal Varian, Robert Vishny and by the editors, Carl Shapiro, Joseph Stiglitz and Timothy Taylor.*

References

Bhattacharya, Sudipto, "Imperfect Information, Dividend Policy and the 'Bird in the Hand' Fallacy," *Bell Journal of Economics*, Spring 1979, *10*, 1, 259–70.

Black, Fischer, and John Cox, "Valuing Corporate Securities: Some Effects of Bond Indenture Provisions." *Journal of Finance*, May 1976, *31*, 2, 351–67.

Black, Fischer, and Myron Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, May-June 1973, *83*, 3, 637–54.

Brooks, Jennifer J. S., "A Proposal to Avert the Revenue Loss from 'Disincorporation,'" *Tax Notes*, July 27, 1987 *36*, 4, 425–428.

Buser, Stephen A., and Patrick J. Hess, "Empirical Determinants of the Relative Yields on Taxable and Tax-exempt Securities," *Journal of Financial Economics*, May 1986, *17*, 335–56.

Canellos, Peter C., "Corporate Tax Integration: By Design or by Default?" *Tax Notes*, June 8, 1987 *35*, 8, 999–1008.

Chamley, Christopher, and Heraklis Polemarchakis, "Assets, General Equilibrium and the Neutrality of Money," *Review of Economic Studies*, January 1984, *51*, 1, 129–38.

Cornell, Bradford, and Kenneth French, "Taxes and the Princing of Stock Index Futures," *Journal of Finance*, June 1983, *38*, 3, 675–94.

Duffie, Darrell, and Wayne Shafer, "Equilibrium and the Role of the Firm in Incomplete Markets," August 1986, Stanford University, manuscript.

Durand, David, "Costs of Debt and Equity Funds for Business: Trends and Problems of Measurement." In *Conference on Research in Business Finance*. New York: National Bureau of Economic Research, 1952, pp. 215–247.

Durand, David, "The Cost of Capital, Corporation Finance and the Theory of Investment: Comment." *American Economic Review*, September 1959, *49*, 4, 639–55.

Fama, Eugene, "Banking in the Theory of Finance," *Journal of Monetary Economics*, January 1980, *6*, 1, 39–57.

Fama, Eugene, "Financial Intermediation and Price Level Control," *Journal of Monetary Economics*, January 1983, *12*, 1, 7–28.

Hirshleifer, Jack, "Investment Decision under Uncertainty: Choice Theoretic Approaches," *Quarterly Journal of Economics*, November 1964, *79*, 509–36.

Hirshleifer, Jack, "Investment Decision under Uncertainty: Applications of the State Preference Approach." *Quarterly Journal of Economics*, May 1966, *80*, 611–17.

Jackson, Thomas H. *The Logic and Limits of Bankruptcy Law*. Cambridge, MA: Harvard University Press, 1986.

Jensen, Michael C., "Takeovers: Their Causes and Consequences," *Journal of Economic Perspectives*, Winter 1988, *2* 21–48.

Kim, E. Han, "Optimal Capital Structure in Miller's Equilibrium." In Bhattacharya, Sudipto, and George Constantinides, eds., *Frontiers of Financial Theory*. Totowa, NJ: Rownan and Littlefield, forthcoming 1988.

Lintner, John, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, February 1965, *47*, 13–37.

Merton, Robert C., "Capital Market Theory and the Pricing of Financial Securities." In Friedman, Benjamin, and Frank Hahn, eds., *Handbook of Monetary Economics*. Amsterdam: North Holland, forthcoming.

Merton, Robert C., "On the Pricing of Corporate Debt: The Risk of Interest Rates," *Journal of Finance*, May 1974, *29*, 3, 449–70.

Miller, Merton H. "The Corporate Income Tax and Corporate Financial Policies." In *Stabilization Policies*, The Commission on Money and Credit. Englewood Cliffs, NJ: Prentice-Hall, 1963, pp. 381–470.

Miller, Merton H., "Debt and Taxes," *Journal of Finance*, May 1977, *32*, 2, 261–75.

Miller, Merton H., "The Informational Content of Dividends." In Dornbusch, Ridiger, Stanley Fischer, and John Bossons, eds., *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*. Cambridge, MA: MIT Press, 1987, pp. 37–58.

Miller, Merton H., and Franco Modigliani, "Dividend Policy, Growth and the Valuation of Shares," *Journal of Business*, October 1961, *34*, 4, 411–33.

Miller, Merton, and Franco Modigliani, "Some Estimates of the Cost of Capital to the Utility Industry, 1954–7," *American Economic Review*, June 1966, *56*, 3, 333–91.

Miller, Merton H., and Myron S. Scholes, "Dividends and Taxes," *Journal of Financial Economics*, December 1978, *6* 4, 333–64.

Modigliani, Franco, "Debt, Dividend Policy, Taxes, Inflation and Market Valuation," *Journal of Finance*, May 1982, *37*, 2, 255–73.

Modigliani, Franco, and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Re-*

view, June 1958, *48*, 3, 261–97.

Modigliani, Franco, and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment: Reply," *American Economic Review*, September 1959, *49*, 4, 655–69.

Modigliani, Franco, and Merton H. Miller, "Corporate Income Taxes and the Cost of Capital: A Correction," *American Economic Review*, June 1963, *53*, 3, 433–43.

Ross, Stephen, "The Determination of Financial Structure: The Incentive Signalling Approach," *Bell Journal of Economics*, Spring 1977, *8*, 1, 23–40.

Ross, Stephen, "Return, Risk and Arbitrage." In Friend, Irwin, and James Bicksler, eds., *Risk and Return in Finance*, Vol. 1 Cambridge, MA: Ballinger, 1976, pp. 189–219.

Rubinstein, Mark, "Derivative Assets Analysis," *Journal of Economic Perspectives*, Fall 1987, *1*, 73–93.

Sargent, Thomas J., and Bruce D. Smith, "The Irrelevance of Government Foreign Exchange Operations," 1986, The Hoover Institution, manuscript.

Sharpe, William F., "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, September 1964, *19*, 425–42.

Spence, Michael, "Job Market Signalling," *Quarterly Journal of Economics*, August 1973, *87*, 3, 355–79.

Stiglitz, Joseph, "A Re-Examination of the Modigliani-Miller Theorem," *American Economic Review*, December 1969, *59*, 5, 784-93.

Stiglitz, Joseph, "On the Irrelevance of Corporate Financial Policy," *American Economic Review*, December 1974, *64*, 6, 851–66.

Stiglitz, Joseph, "Information and Capital Markets." In Sharpe, William F., and Cathryn Cootner, eds., *Financial Economics: Essays in Honor of Paul Cootner*. Englewood Cliffs, NJ: Prentice-Hall, 1982, pp. 118–58.

Stoll, Hans R., "The Relationship Between Put and Call Option Prices," *Journal of Finance*, December 1969, *24*, 5, 801–24.

Wallace, Neil, "A Modigliani-Miller Theorem for Open Market Operations," *American Economic Review*, June 1981, *71*, 5, 267–74.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 121–126

Why Financial Structure Matters

Joseph E. Stiglitz

The 1958 paper by Franco Modigliani and Merton Miller has been justly hailed as a landmark in the modern theory of finance. What has not been sufficiently emphasized is the importance of the paper to the development of economic theory and practice. Indeed, it is ironic that a paper which purportedly established that one need not pay any attention to financial structure—that financial structure was irrelevant—should have focused economists' attention on finance. Merton Miller gives what must be part of the explanation in his preceding paper in this issue: by providing conditions where financial policy was irrelevant, conditions which were close to the assumptions used by most conventional economists, the paper forced a reexamination of those standard assumptions. That reexamination is still going on.

The MM results have such great intellectual power and appeal that their direct effect has sometimes been to lead economists astray. This result may perhaps be seen most forcefully in the work in investment theory. In the major study done prior to MM, that of Meyer and Kuh (1959), financial variables (like profitability) were identified as having important independent effects on investment. But the paradigm that became dominant following the publication of MM, most forcefully articulated by Dale Jorgenson and his associates, argues that in the absence of taxation, financial structure (for instance, the magnitude of the firm's equity base) or cash flow would make no difference to the level of investment. Theory drove the econometrics: financial structure variables were excluded because "economic theory"—that is, Modigliani and Miller—said they should be excluded. Only recently, as a developing and substantial body of economic theory says once again that such variables should be included, have econometricians included profit variables again in their specifications

■ *Joseph E. Stiglitz is Professor of Economics, Stanford University, Stanford, California.*

of the causes of investment. And lo and behold, they appear to be significant! One of the lessons to be learned here is that economists should be careful in being overly dogmatic in the theoretical strictures we impose on our econometric models. Though it may be going too far to advocate "letting the data speak for itself" without any theoretical guidance, economists should at least be attentive to those whispers which the data occasionally emit.

Again ironically, some of the most productive responses to the MM results have come from those who did not feel able to accept the conclusion that financial policy is irrelevant. The MM results forced these skeptics to identify which of the assumptions underlying the MM theorem should be modified or rejected. The attention of economists during the past 30 years has focused on four assumptions underlying the model: first, that firms can be identified by "risk class;" second, that individual borrowing can substitute for firm borrowing; third, that investors have full information about the returns of the firm; and fourth—the importance of which MM themselves recognized—is that there are no taxes, or at least tax policy does not treat debt and equity differentially. The question has been not so much whether these assumptions are "realistic," but whether, or under what circumstances, altering these assumptions leads to situations where financial structure does indeed matter.

The Risk Class Assumption

The assumption that each firm could be sorted into a "risk class" was a simplification which turned out to be enormously useful in other areas. But strangely enough, it turned out that the risk class assumption—and the associated arbitrage argument—could be completely dispensed with and the validity of a more general proposition on the irrelevance of financial structure, within a general equilibrium context, could easily be established, provided only that firms do not issue so much debt that they incur a positive probability of bankruptcy (Stiglitz, 1969). The restated theorem asserted simply that, if there is an equilibrium with a firm having a particular debt-equity ratio and market value, there exists another equilibrium with the firm having any other debt-equity ratio, and in that new equilibrium the firm has the same market value as it did in the original equilibrium.[1]

This result shifted the focus of attention from the idea of the risk class to the importance of bankruptcy. The market does seem to be concerned about bankruptcy; after all, interest rates charged even the best of firms exceed by a considerable amount the rates paid by the government. Yet the fact that bankruptcy is important, in this sense, does not necessarily invalidate the relevance of the revised MM theorem. Economists have tried to spell out models where a chance of bankruptcy exists, but financial policy remains irrelevant. For instance, if the returns of all firms were

---

[1] In a subsequent paper (Stiglitz, 1974), it was shown that not only did the firm's choice of debt-equity ratio not matter, but neither did its dividend payout ratio, the maturity structure of its debt, nor any other aspect of its financial policy.

(jointly) normally distributed and all individuals had quadratic, constant absolute risk aversion, or constant relative risk averse utility functions, or if there were a sufficient number of securities so that the value of income in each state of nature (each outcome) was determined independently of what the firm did, then the financial irrelevance proposition would still be valid.

But none of these "special" models has gained broad acceptance, because each of them has some unrealistic implications. For instance, what firm managers mean by risk does not accord well with how the models predict managers should use the term; in these models, the only aspect of risk which is important is how an individual firm correlates with the market return, but firm managers also care about the variance of the return. What is at issue is not just a matter of semantics: the information that firms gather for decision making is based on their view of the appropriate concept of risk.[2] Moreover, most of these models also argue that the value of the firm is proportional to its scale, or equivalently that firms do not face downward sloping demand curves for their shares. Again, several studies have at least called into question the empirical validity of that hypothesis.

Homemade Leverage

A second assumption which entered both the original MM presentation and subsequent generalizations was that individual borrowing, or homemade leverage, was a perfect substitute for firm borrowing. This assumption seemed to contradict the obvious fact that many individuals were constrained in the amount of credit they could get, and in any case could not borrow at the same terms that firms could. Much of the subsequent finance literature ignored the possibility of individuals being credit-constrained, not because of the empirical evidence showing that such constraints did not exist, but again because of theoretical considerations: in a perfect competitive capital market, such constraints simply could not exist. There were voices in the wilderness—James Tobin comes to mind—who contended that credit constraints were important, but they were not listened to until theory provided a rationale for such constraints.

Full Information about Returns

It is often the unstated rather than the stated assumptions of a model which are critical, and so it is in the case of the MM theorems. A critical unstated assumption is that all market participants have full and equal information concerning the returns to the firm. However, the existence of asymmetric information gives rise to two problems:

---

[2] There are other problems, particularly with the mean-variance models. For example, they "predict" that all individuals purchase the same mix of risky securities and that individuals should hold widely diversified portfolios, two predictions that seem to be contradicted by the facts.

current owners may wish to convince potential borrowers that the firm is worth more; and managers can take actions which affect the returns to those who provide capital. In the simple early models where everyone has full information, debt and equity differed only in the sort of risk they involved. In the world of imperfect information they differ in a number of other respects: since the amount that suppliers of debt receive is fixed, except in states of bankruptcy, debt does not require monitoring total returns; all that is required is monitoring actions which affect bankruptcy, and monitoring income in bankruptcy states. The fact that equity is widely dispersed, while debt (partly because of its risk characteristics) can be more closely held means that the free rider problems associated with ensuring that managers take actions in the interests of equity holders are greater than the corresponding problems of ensuring that managers take actions in the interests of debt holders; as a result, it has been argued that debt holders may exercise more effective control over capital than equity holders.

Managerial incentives are markedly different with debt and equity contracts. Debt imposes a threat of bankruptcy on managers, a threat which advocates of what Robert Hall has called the "backs-to-the-wall" approach to corporate finance believe will have a positive effect on managerial incentives. By the same logic, debt and equity contracts have very different effects on managerial risk taking.

Because of these and other differences between debt and equity, a firm's choice of financial structure may convey information about the firm's prospects to potential lenders and purchasers of equity. Theoretical models have been constructed to show why equity issues may be interpreted as a negative signal, and empirical evidence confirms their depressing effect on price.[3] Owners who have inside information and know that the market has overvalued their shares are more willing to sell shares than those who know that the market has undervalued their shares. Managers who know that their true net worth position is favorable may be willing to signal that through larger cash distributions to shareholders. Managers who know that there is little prospect of bad outcomes may be willing to signal that knowledge by undertaking more debt.

These considerations of adverse selection and moral hazard have been used to explain why firms may be credit-rationed (Stiglitz and Weiss, 1981, 1983; Keeton, 1978) or equity-rationed. But these forms of financial constraints are but extreme examples, demonstrating clearly that the financial structure of firms can make a difference. Theories of optimal financial structure can be derived, with the optimal structure depending on the nature of the information problems being faced.

Tax Differentials

A final set of circumstances in which financial structure may matter relates to the existence of taxes. Debt and equity are treated by tax authorities in markedly different ways. While the fact that debt payments are deductible under the corporation tax might seem to give debt an advantage over equity, one needs to look at the total tax consequences—at the combined effects at the individual and corporate levels. Moreover, the tax laws impose taxes on changes in financial structure. In other words, even if two firms had identical earnings, but one had a higher debt and lower total tax liabilities, it might still not pay for the other firm to change its financial structure. To increase its debt-equity ratio, after all, the second firm must pay out money to its equity holders (borrowing to pay dividends or buy back shares), and there is thus a cost of transition—the tax on the dividends, or the capital gains tax on the share repurchase—which must be set against the steady state gains.

---

[3] For theoretical discussions, see Leland and Pyle (1977), Ross (1977), Stiglitz (1982), Greenwald, Stiglitz, and Weiss (1984), and Meyers and Majluf (1984). For a discussion of empirical evidence, see Asquith and Mullins (1983).

What is required is a full, dynamic analysis[4] of firm financial structure, incorporating both corporate and individual taxation, of the kind I presented in my 1973 paper. There, I showed that for tax rates of the kind that had prevailed up to that time, while debt had a slight steady state advantage over equity, this advantage was not so great as to make financial restructuring desirable. In general, it did not pay firms that had small debt-equity ratios, as a result of a series of high returns on previous investments, to increase their debt-equity ratio. If their demand for investment exceeded their retained earnings, it was optimal for them to raise the additional funds required through debt. The fact that there is, in effect, a tax on financial restructuring means that there is a large historical element in observed capital structures. And it explains why changes in tax laws, such as the 1986 Tax Reform, seem to lead to, at most, limited financial restructuring by most firms.

There remain puzzles. Though one can easily construct a theory in which it pays firms to distribute funds to shareholders to signal the worth of the firm, it is always desirable for firms to do so by repurchasing shares or by cash mergers and acquisitions, rather than by dividend payments. The suggestion that, were all firms to repurchase shares, the IRS might change the rules[5] is not an explanation for why any single firm issues dividends; and even were it to tax share repurchases, there are other means by which funds may be distributed from the corporate to the household sector and receive favorable tax treatment. Since this paradoxical question of why dividends should be paid at all was first discussed (for example, in my 1973 paper), firms have in fact seemed slowly to learn this lesson to the point where more funds are now distributed in these forms than as dividends. (A paper forthcoming in this journal by John Shoven and Laurie Simon Bagwell documents and discusses this trend.) Yet why were firms so slow to learn this costly lesson, and why have so many firms yet to learn it at all?

[4] Such a dynamic analysis must take explicit account of the asymmetries in the tax code. For example, positive cash flows from the corporate sector to the household are taxed, but there is no negative tax on negative cash flows.

[5] Under current tax law, sales of shares back to the firm are treated just like any other sale of shares; the individual pays a capital gains tax on the difference between the sale price and the price he originally paid for the shares. There are a variety of ways by which the tax authorities could discourage share repurchases. Prior to 1986, they could have subjected share repurchases to full taxation, rather than extending the favorable capital gains treatment. They could subject the total purchase price to taxation, not allowing a deduction for the original purchase price.

The fact that the optimal financial structure depends on tax rates, and that different investors face markedly different tax rates means, of course, that different firms should have different clientele. If those versions of the capital valuation model which assert the existence of a large number of similar firms were correct, then there would be some firms which pursued an all debt policy, others an all equity policy, and still others a "historical" policy of the kind described above; within any risk class (to use the MM concept) firms in each category would have distinctly different owners. Though firms may differ somewhat in their clientele, they do not differ in the marked way predicted by the theory.

Modigliani and Miller, in their brilliant papers, have set forth a research agenda which will occupy economists for decades to come. The direction which Modigliani and his co-authors suggest in their recent work, a systematic exploration of market "irrationalities," seems among the most fruitful for enhancing our understanding of these hitherto unexplained quandaries.

References

Asquith, P. and Millins, D. W., "Equity Issues and Stock Price Dilution," Harvard Business School, unpublished paper, 1983.

Greenwald, B., Stiglitz, J. E., and Weiss, A. "Informational Imperfections in the Capital markets and Macro-Economic Fluctuations," *American Economic Review*, May 1984, *74:1*, 194–99.

Keeton, W., *Equilibrium Credit Rationing*, New York: Garland Press, 1978.

Leland, H. E., and D. H. Pyle, "Informational Asymmetries, Financial Structure, and Financial Intermediation," *Journal of Finance*, 1977, *32*, 371–87.

Mayer, J., and E. Kuh, *The Investment Decision; and Empirical Study*. Cambridge: Harvard University Press, 1959.

Myers, S. C., and N. S. Majluf, "Corporate Financing and Investment Decisions when Firms have Information that Investors Do Not," *Journal of Financial Economics*, 1984, *11*, 187–221.

Ross, S. A., "The Determination of Financial Structure: The Incentive Signalling Approach," *Bell Journal of Economics*, 1977, *8*, 23–40.

Stiglitz, J. E., "A Re-Examination of the Modigliani-Miller Theorem", *American Economic Review*, December 1969, *59:5*, 784–793.

Stiglitz, J. E., "Corporate Financial Policy and the Cost of Capital," *Journal of Public Economics*, February 1973, *2*, 1–34.

Stiglitz, J. E., "On the Irrelevance of Corporate Financial Policy," *American Economic Review*, December 1974, *64*, 851–866.

Stiglitz, J. E., "Information and Capital Markets," In Cootner, C., and W. Sharpe, eds., Financial Economics: Essays in Honor of Paul Cootner." Englewood Cliffs, NJ: Prentice-Hall, 1982, 118–58.

Stiglitz, J. E., and A. Weiss, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71:3*, 393–410.

Stiglitz, J. E., and A. Weiss, "Incentive Effects of Termination: Applications to the Credit and Labor Markets," *American Economic Review*, December 1983, *73*, 912–927.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 127–133

Comment on the Modigliani-Miller Propositions

Stephen A. Ross

What a treat it is to have the opportunity to read Merton Miller's ruminations on the MM Propositions. If the view of the progress of science that interprets it as one of changing paradigms has merit, then surely the work of Miller and Modigliani provides a laboratory example of a violently shifted paradigm.

Through Miller's eyes we can recapture some of the flavor of how dramatic that change was. We are now so accustomed to the acceptance of the new paradigm that the older view that capital structure did matter has about it the flavor of phlogiston. Not only does it seem wrong, it is difficult to believe that sensible folk could have held such beliefs. In fact, though, as with phlogiston, not merely did esteemed scholars subscribe to the view that capital structure did matter, the stubborn debate that followed the first publication of the MM papers attests to their devotion.

Because economists now do look at finance through the eyes of MM, one of the contributions of a retrospective appraisal is that it affords us the opportunity to explore what, if anything, there was of value in the older view that we might have lost in our zeal to replace it. This is what Miller emphasizes when he says that "showing what *doesn't* matter can also show, by implication, what *does*."

As their work has been recast in the shifting landscape of ever more modern theory, there is a danger that the original analysis will become discarded as "crude" or "naive." This would be a terrible mistake; the older analysis was designed less to verify some mathematical truism than to capture a live and elusive scientific intuition. Since the newer theory rests on a somewhat different base of intuitions, some of the older insights embedded in the original proofs and analysis may be lost. Miller's piece provides an opportunity to do some excavating.

■ *Stephen A. Ross is Sterling Professor of Economics and Finance, Yale School of Organization and Management, Yale University, New Haven, Connecticut.*

A comment on Miller and Modigliani's work can easily become a discussion of any facet of modern finance. To avoid overlapping unnecessarily with the other comments, I will focus my dig in the two areas of arbitrage and taxation.

Arbitrage Analysis

Since the original MM analysis economists have learned an enormous amount about the role played by no arbitrage conditions in financial markets. We have discarded the risk class based arbitrage arguments and no longer have a need for the perfectly correlated companion firm that MM used. Now we merely invoke the result that the absence of arbitrage implies the existence of a linear pricing rule (or, equivalently, an expectation) that can be used to value all assets as a function of their payouts (or their expected payouts using risk adjusted probabilities).[1]

To illustrate the new analysis, consider the typical firm whose total value is equal to the sum of the values of its debt and its equity. The values of debt and equity, in turn, are based on the cash flows that are expected to go to each. Since the cash flow of the firm must go either to debt or to equity and since linearity assures us that the sum of their values is the value of the sum of their cash flows, it follows that the total value of the firm depends simply on the total cash flow of the firm, regardless of the debt-equity mix chosen by that particular firm.

But have we really learned that much? And if we have, what do we now know about the original MM proofs? The original arguments went something like this. Imagine two firms in the same risk class; in particular, assume that they have identical cash flows. Suppose that one is capitalized with equity alone and that the other has both debt and equity outstanding. Lastly, suppose that the levered firm's value is less than that of the unlevered firm.

Purchasing an equal percentage share of the levered firm's debt and equity would then cost less than the same percentage share of the all equity firm, but it would entitle an outside investor to exactly the same cash flow. Such an arbitrage possibility would raise the price of the levered firm's equity and lower that of the unlevered firm until the two firms had the same value.

Conversely, if the levered firm were relatively overvalued, then by combining the unlevered firm's equity with borrowing—that is, by creating "homemade leverage" —the investor could duplicate the return on the levered firm's equity at lower cost. Of course, this scenario raises questions about the ability of investors to borrow on their own accounts as cheaply as the levered firm.

What happens to the features of these arguments when we bring in the deux machina of modern arbitrage pricing analysis? For one thing, we do away with the need for a replicating firm in the same risk class. The risk class argument is a way of making the point that in cross section the firm will not alter its value by altering its capitalization. It is a subtle device. When a firm changes its capitalization this has potential effects on the entire financial equilibrium. This, in turn, clouds the analysis immeasurably by taking it into the realm of general equilibrium where the action has rippling effects on prices and actions throughout the economy. The replicating firm is a device for eliminating this problem and confining the argument to a partial equilibrium. As long as two identical firms coexist, they cannot differ in value.

---

[1] Varian (1987) and Dybvig and Ross (1987) provide introductions to the modern no-arbitrage analysis.

The modern arbitrage analysis, while it hides some of the machinery of arbitrage, repays us by making this point more explicit. As long as changing the firm's capitalization does not alter the pricing of risky cash flows in the economy as a whole, then changing the firm's capitalization cannot alter its value.

However, strictly speaking, this is not what the MM arguments prove. Rather, they only show that within their risk class *relative* values are unaffected by capitalization. There is nothing in their arguments that would preclude the firm from changing its capitalization and raising or lowering the value of *both* itself and the unlevered firm. In other words, strictly speaking, the original analysis really only works when the unlevered firm's value is pegged. But this is just an implication of the assumption that the economy wide pricing of cash flow streams is unaltered.

This takes us back full circle to the original Marshallian and neoclassical analysis of perfectly competitive markets with constant returns to scale. The supply curve is perfectly elastic in such a market because the expansion of any one firm's output can be accommodated by an equivalent contraction in the output of the other firms. Since equilibrium profits are zero the real equilibrium is unaltered.

Similarly, when one firm raises its debt-equity ratio, its value is unaltered because the pattern of risky cash flows that it is supplying to the capital market can be duplicated elsewhere in the economy at no cost. If this picture of the world seems a bit farfetched, one might be tempted to succumb to pragmatism and simply rely on the intuition that the change would, at most, have a small effect, but there is merit in the pristine certainty of the neoclassical world.

In either case, the richness of substitutes for the cash flows offered by any one firm now takes the place of the identical firms in the original MM risk classes. Depending on the pricing rule for the economy, this is a truly meaningful extension. For example, unpriced idiosyncratic risk can imply that even if the random characteristics of a firm's cash flows differ dramatically from those of other firms in the economy, nevertheless the idiosyncratic component of the cash flow will be valueless and will not affect the substitutability or the pricing of the firm (although it will influence the *relative* pricing of the debt and the equity).

The argument is the familiar one from insurance. Two similarly situated individuals will pay the same for life insurance not because their deaths will be perfectly synchronized but, rather, because they are independent. As a consequence, at the margin they contribute nothing to the risk of a large pool and can receive "fair" insurance. The same pooling of independent risks occurs in the financial markets and two securities can be perfect substitutes in large portfolios even if they are quite different. Indeed, we can now replace the two identical firms in the MM arbitrage argument with two firms with the same value and with completely *uncorrelated* cash flows.

The Concept of a Risk Class

This process of understanding how the economy allows investors to duplicate the risky return of any individual company should be understood as an expansion of the original MM notion of a risk class. The "risk class" played an important role in the original arbitrage analysis, as Miller explains, but it has subsequently passed from favor. However, I think that it might be time for a revival of a modern perspective on the older views. This is particularly so given the sorry empirical state of our asset pricing theories.

As I have argued, Miller adopts too narrow a view of a risk class when he refers to it as a "spanning set" consisting of firms such that "the uncertain, underlying future cash flow streams of the firms within each class could be assumed perfectly correlated, and hence substitutes." To some extent this limited interpretation may be responsible for the atrophied interest in the concept.

Two cash flow streams that are perfectly correlated differ only by a sure cash flow and a difference in scaling. From the perspective of the pricing of risky assets they are essentially identical. This is the same identity that underlies modern option pricing theory, but the intuitive appeal of the notion of a risk class is broader than the observation that two identical cash flows must have the same value.

Presumably the concept has merit because a risk class can be defined under less stringent requirements; that is, two cash flow streams need not be perfectly correlated to be meaningfully placed in the same risk class. If an implication of being in the same risk class is that securities have the same cost of capital, then, since the cost of capital depends on the systematic risk of a company, we have seen that two firms with completely *uncorrelated* risks, risks that are independent of the major systematic risks in the economy, can be placed in the same risk class.

The important possibility opened up by the concept of a risk class is that of arguing on a priori grounds that the structural and financial characteristics of two firms are such that they would respond similarly to exogenous economic shocks and, as such, belong in the same risk class. If this could be done successfully in a cross section analysis of firms, then it would free the question of the determination of the cost of capital from an allegiance to one or another of the competing asset pricing models.

Whether or not such further research on the concept of a risk class will be fruitful has direct implications for empirical work in the MM theory, and to the extent to which such research leads to new insights on the implications of the absence of arbitrage, it bears directly on the theory itself. Miller rightfully considers what might be called the fine structure of the arbitrage arguments. In response to whether individuals could actually implement the homemade leverage he points out the inessential nature of the limited liability assumption and he notes that corporations as well as individuals hold securities and can arbitrage. He also observes the seeming obstacles to arbitrage posed by the special features of common stock such as voting rights. In a sense these make for a different sort of arbitrage and, presumably, a more difficult sort.

But, focusing on these shortcomings of arbitrage arguments seems to me to ignore those very changes in the financial markets to which Miller calls our attention. While the range of corporate instruments is not really much more innovative today than it was even a century ago, the range and the volume of trading in noncorporate securities—including securities such as call options that, while not issued by corporations, are derivative to them—has increased exponentially.

This means that the major force of arbitrage appears less at the firm level than it does within the capital market itself. The number of close substitutes for any security is now so deep and rich that any corporate security can now be priced by arbitrage. To put it in terms of our previous discussion, we need only find the security's risk class.

Arbitrage and Taxation

Taxation is where the real world meets the theory. At first blush, the corporate deduction for interest seemed to doom the MM analysis to an academic curiosity, since it appeared to give debt an obvious cost advantage over equity. It was the genius of Miller to see that the arbitrage analysis went much too deeply into the workings of the financial markets and the economy as a whole to be easily unseated by even something as powerful as the tax code.

Although the original MM papers dealt with this issue, it surely reached maturity in Miller's (1977) seminal paper on debt and taxes, a paper which should be thought of as a companion piece to the original MM papers. (See, also, previous work on these matters by King (1974) and Stiglitz (1973).) Miller acknowledged that the corporation would enjoy higher total after-tax income by increasing its debt-equity ratio, and that this additional income would also result in a higher gross combined payout to stockholders and bondholders. But his tactic did not mean that the value of the firm would have to increase. As debt is substituted for equity, the proportion of the payout in the form of interest on the debt rises relative to the payouts of dividends and capital gains on equity. To the extent that investors are relatively more heavily taxed on interest payments than on equity returns, this cuts into the corporate advantage to the issuance of debt.

From the perspective of the corporation, debt now costs more relative to equity than it did, with the wedge between them equalling the marginal investor tax advantage of equity over debt. Miller further showed that with a rising marginal tax rate for investors, a sensible equilibrium would be one in which the corporate cost of equity was below that of debt by precisely the marginal corporate tax advantage of debt.

Of course, as Miller points out again in his comment, this logic does not imply that the aggregate issuance of corporate debt need be indeterminate. In fact, his equilibrium provides a unique solution for the proportion of corporate debt outstanding. Once this equilibrium is reached no corporation would see a tax advantage to the issuance of debt and the original irrelevance proposition would be restored. As Miller also noted, though, there would not be an internal equilibrium if the wedge between the personal tax on interest income and that on capital gains was too small relative to the corporate tax rate. In fact, back of the envelope calculations, particularly with the recent changes in the tax code, reveal that the personal advantage to receiving capital gains is too low to compensate for the corporate tax.[2]

Without the offsetting effects from the personal tax code, the only brake to the expansion of firm debt would be the recontracting costs implicit in bankruptcy. Since the costs of bankruptcy seem relatively small in comparison with the tax advantages of corporate debt, it is difficult to explain the relatively conservative debt policies of many corporations. This is exacerbated by the possibility of income bonds and other devices for minimizing the bankruptcy potential while retaining the tax advantage. Leveraged buyouts and other vehicles for creating explicitly risky debt cum control rights are a more recent approach to the same problem.

These insights put academic economists in the awkward position of finding corporate America unwilling to take some rather obvious actions to maximize value. As Miller notes, this is not very comforting for advocates of positive economics. Perhaps to ease their burden, at the end of his comment Miller alludes to an infinity of ways by which business can avoid or lessen the impact of the double tax hit on corporate income. Ross (1985) explores some of these possibilities, and it is here that I find the true lesson of the intuition behind the MM propositions and the absence of arbitrage.

I would like to conclude by offering a modest if extreme version of this intuition. The mathematician, Godel, argued that no finite set of axioms for mathematics could ever be sufficient in the sense that there would always remain undecidable propositions that could be added or not without affecting the integrity of the original system. It seems to me that the time has now come to marry Godel and the IRS with MM presiding at the ceremony. To wit, we now have an explanation for the secular expansion in the size of the tax code:

> *Godel and Miller's Tax Proposition*: No finite and feasible system of business taxation can collect positive revenues.

Feasible just means that no one is required to pay more taxes than they have. In a world of less than perfect information and expectations the tax code grows as the government tries to plug the loopholes that continually arise out of the inability to specify all of the possible contingencies and the groping by business to avoid taxation. (In the process the corporate tax does seem to raise a good piece of change.)

Proving this proposition is more difficult than stating it, but perhaps the modern combination of the no arbitrage analysis with the tax code provides a clue. To begin with, we know from the no arbitrage analysis that the value of a firm's income is a function not only of the total cash flow but also of the forms in which it comes packaged, including depreciation offsets, capital gains, and so on (see Ross, 1987). Given a price for each of these, the firm will package itself so as to maximize its value.

---

[2] Even the personal tax effect is not determinate and, depending on the code and other features of the economy, very different equilibria are possible. In some, the corporate gains from leverage dominate and maximal debt is optimal; in others, the gains are actually losses and firms retreat to all equity.

Ignoring organizational costs, firms will then combine, split and reorganize into entities with different tax rules in such a way as to minimize the tax.

While our conjecture doesn't quite merit the usual Q.E.D., it does serve to illustrate a proposition that surely does. Namely, that economists owe a debt to Miller and Modigliani that is at least as great as that which mathematicians and philosophers owe to Godel.[3]

■ *The author is grateful to the editors, Carl Shapiro, Joseph Stiglitz, and Timothy Taylor, for their helpful comments.*

---

[3] Or is it equity? Well, it doesn't really matter, does it?

References

King, Mervyn A., *Public policy and the Corporation.* London: Chapman & Hall, 1977.

Miller, Merton H., "Debt and Taxes," *Journal of Finance*, May 1977, *32:2*, 261–75.

Ross, Stephen A., "Debt and Taxes and Uncertainty," *Journal of Finance*, July 1985, *60:3*, 637–57.

Ross, Stephen A., "Arbitrage and Martingales with Taxation," *Journal of Political Economy*, April 1987, *95:2*, 371–92.

Dybvig Philip H., and Stephen A. Ross, "Arbitrage." In Eatwell J., M. Milgate and P. Newman, eds. *New Palgrave, A Dictionary of Economics.* London: The MacMillan Press, Ltd. 1987, *1.* 100–106.

Stiglitz, Joseph E., "Taxation, Corporate Financial Policy and the Cost of capital," *Journal of Public Economics*, February 1973, *2:1*, 1–34.

Varian, Hal, "The Arbitrage Principle in Financial Economics," *The Journal of Economic Perspectives*, Fall 1987, *1:2*, 55–72

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 135-147

Corporate Finance and the Legacy of Miller and Modigliani

Sudipto Bhattacharya

The influence of the Modigliani-Miller (1958) propositions on capital structure and the Miller-Modigliani (1961) theses on dividend policy permeates almost all aspects of financial economics to this day. The implications of their methodological innovations for later work on option pricing, on the constrained optimality of stock markets, and in public finance and macroeconomics have been noted elsewhere in this symposium. In this commentary, I shall focus on the influence of Miller's and Modigliani's contributions on a couple of key areas in corporation finance, and review research progress by later contributors.

My choice of these topics has been motivated by two distinct criteria: their central importance in understanding corporate financial policy and equity valuation, and their preeminence in later contributions by Miller (1977) and Modigliani (1982) in their presidential lectures to the American Finance Association, as well as in Miller (1987) and Modigliani and Cohn (1979), for example. Broadly speaking, these two themes can be summarized as: (A) integrated tax- and information-related considerations in capital structure and dividend policy choices; and (B) the impact of inflation and nominally denominated debt contracts on the valuation of corporate equity.

Debt and Dividend Policy Choices

The key methodological contributions of the MM (1958, 1961, 1966) papers were, in my opinion, those of: (a) introduction of the risk-class (set of payoff patterns mutually replicable through trading) notion; (b) consideration of investor arbitrage

■ *Sudipto Bhattacharya is Professor of Finance, School of Business Administration, University of Michigan, Ann Arbor, Michigan.*

(homemade leverage) in pricing securities; (c) initiation of integrated after-tax analyses of dividend and debt supply policies of firms; (d) consideration of empirical evidence and introduction of "respectable" econometric methods in corporate finance; and (e) planting seeds for the development of economic modeling of unexplained phenomena, such as the "informational content" of dividends for stock prices. In this section, I shall focus in particular on topics (c) and (d), and developments along the lines of topic (e). Extensions of the MM methodology on (a) and (b), to considerations of alternative equilibria and the endogeneity of risk classes of equities in infinite-horizon economies, will be germane to the discussion of the impact of inflation and leverage on equity valuation in the next section.

From MM (1963) we knew that, based on the differential tax deductibility of interest costs at the corporate level, the addition of D dollars in market value of debt should add a tax-shield value ($T_c \times D$) to firm value, where T_c is the corporate tax rate, if the following assumptions hold. First, the debt is a perpetuity; second, interest is effectively tax-deductible (earnings are nonnegative) whenever it is paid; third, there are no other costs of debt, and no information about earnings prospects is communicated by the use of debt. In Miller (1977), this calculation was modified to reflect the tax advantage of equity at the personal level, owing to lower and postponable taxation of capital gains (see the equation in his current article). This analysis (and Stiglitz, 1973) raised the possibility that choices of corporate leverage could be determinate (by investors' demands) only at the aggregate level, with indifference for valuation at the level of the firm.

Miller's work also focused on the tax consequences of that other major financial policy variable, the payment of dividends. In an often-cited article, Miller and Scholes (1978) suggested that the (pre-1986) personal tax disadvantage of dividends relative to capital gains could be obviated by investors' portfolio strategies. Specifically, taxable investors would borrow to generate interest deductions equal to their dividend income, and invest the borrowing proceeds in tax-exempt means of accumulation such as cash-value life insurance (subject to legal limitations to pursuing such schemes; see also Stiglitz, 1973). Together, Miller's papers advanced the notion of firm-level *indifference* to debt and dividend policies for valuation, as a candidate solution for market equilibrium.

Empirical evidence has, by and large, gone against the above parsimonious view of corporate financial policy. Studies of the marginal bondholder tax rate across tax-exempt (municipal) and taxable bonds, exemplified by the recent paper of Buser and Hess (1986), suggest the following. The tax rate T_{PB} of the marginal bondholder indifferent between taxable and tax-exempt bonds is on average significantly below the corporate tax rate T_C. The differential between these two rates, and the resulting positive marginal tax-shield value of corporate debt (as measured by the equation in Miller's article) suggest a major role for other offsetting costs of debt in determining optimal leverage choices. (See Stiglitz (1985), as well as Kim (1988), for models of such costs.) These costs include (a) loss or postponement of interest tax shields due to future bankruptcy, (b) the present value of prospective bankruptcy-related costs, and (c) the "agency costs" of managerial behavior—such as under-investment and overly risky strategy choices—that can occur when levered equity holders have limited liability.[1] These costs, and thus the required offsetting tax shield of debt, are likely to be higher when corporate earnings are depressed, with some temporal persistence, since that increases the risk of default at a given level of debt. The empirical evidence of Buser and Hess on the cyclical behavior of T_{PB} indirectly confirms this prediction of theories that model interior choices of debt-equity ratios, choices that arise from the tradeoff between the net positive tax-shield value versus other "deadweight" costs of debt.

Evidence on the impact of dividend policy on equity valuation is also, on balance, against the Miller-Scholes (1978) solution. However, it is fair to admit that there is some controversy over the interpretation of this evidence (for example, Litzenberger and Ramaswamy, 1980, vs. Miller and Scholes, 1982). Much of this controversy has to do with the interpretation of the observed impact that dividend yields have on the risk-adjusted returns on equities. It has been known for a long time that unanticipated (relative to trend) announcements of dividend increases are associated with contemporaneous increases in stock prices, a linkage often termed the informational content of dividends (Miller, 1987). The essence of the difference between Miller-Scholes and Litzenberger-Ramaswamy has to do with determining the "right" proxies for anticipated dividend yields, so that higher risk-adjusted average returns on high-dividend stocks can be interpreted as a tax-induced effect that equates returns net of personal taxes.

Based on studies of stocks on the New York Stock Exchange (NYSE), Litzenberger and Ramaswamy (1980) and others (for example, Keim, 1985, in non-January months) conclude that an extra percentage of dividend yield causes investors to demand roughly a 0.2 percent increase in total pre-tax equity returns. Miller and Scholes (1982), using different lagged dividend yield measures, arrive at a much smaller number close to zero. However, there is a further wrinkle. Since investors in different personal tax brackets (e.g., pension funds versus individual investors), should hold different portfolios of high- and low-dividend stocks, the impact of dividends on required returns should be different in separate yield categories. Both Litzenberger-Ramaswamy and Miller-Scholes, using very different dividend-yield measures, find this to be the case. As Modigliani (1982) notes, this similarity alone creates some presumption in favor of the contention that the higher required returns on high-dividend stocks are due to tax effects. On balance, the evidence suggests that, prior to the 1986 Tax Reform Act, firms would have reduced the required return on their equities by reducing their dividend payout.

For large changes in dividend policy though, other valuation tradeoffs can come into being. The seminal analysis of MM (1961) assumed that lower dividends (greater retention) would be matched by lower new equity issue, holding the levels of investment and debt constant. For many firms outside the regulated utilities industries

[1] The work of Stiglitz and his coauthors (see Stiglitz, 1985, for a summary), on these incentive effects of capital structure, provides important qualifications to the MM (1958) leverage indifference proposition. The use of income bonds, with interest paid given sufficient income, or collateralized junior debt (Bhattacharya, 1988), can ameliorate some of these agency costs of debt arising from limited liability.

—which can pass through their cost of capital to the rate base—new equity issues are few and far between, and the relevant tradeoff holding investment constant might be that of greater retention leading to a lower level of debt issue. If, as some find plausible (Modigliani, 1982), the *net* positive tax shield per unit of debt was around 20 percent, and this was comparable to the marginal tax disadvantage of a dollar of dividends for valuation, then this debt-dividend tax tradeoff might have been a matter of indifference. Further work on this issue is warranted, although the 1986 Tax Reform Act has drastically reduced the personal tax advantage of capital gains relative to ordinary (dividend or interest) income, and thus also increased the tax shield value of corporate debt.

What about the analytical assumption, recommended by MM (1961), that investment be held constant when analyzing dividend policy choices? Several recent studies of the dividend payout "puzzle" (Auerbach, 1984; DeAngelo, 1987), assume that real investments are made by firms with retained earnings which would *not* be made with new equity issues, until the point at which the alternative of paying taxable dividends to equity dominates. However, the feasibility of firms using retained earnings to purchase marketed securities voluntarily held by households, together with the low taxation of intercorporate dividends, casts doubt on this phenomenon as being the major determinant of corporate investment and dividend payout policies.

An alternative set of models, grounded in *asymmetric information* between corporate insiders and outsiders (future shareholders), seeks to simultaneously explain both the payout and informational content aspects of dividends.[2] One set of examples, including Bhattacharya (1979), John and Williams (1985), and Miller and Rock (1985), focuses on circumstances in which: (a) current earnings of firms are asymmetrically known by insiders who serve current shareholders' interests; (b) shareholders who need to sell, currently or in the future, care about the firm's valuation for their overall returns; and (c) dividends paid have deadweight costs such as personal taxes, corporate transaction costs of refinancing cash flow shortfalls, or costs of under-investment. In these circumstances, it can be shown that firms may use dividends as a costly *signal* of the Spence (1973) type, to differentiate each other and communicate higher (anticipated) earnings to the stock market with higher dividends. Given knowledge of the intertemporal stochastic process of economic earnings, the market then arrives at informationally accurate current valuations of firms that are increasing in the level of dividend payout. The reason dividends work as a signal is that the expected marginal cost of *some* component of the deadweight cost of dividends is negatively related to (expected) earnings, so that only firms with higher earnings prospects are induced to pay higher dividends in a separating equilibrium.[3]

[2] Almost any conceivable resolution of the dividend payout puzzle has to assume that a systematic policy of repurchasing equity would result in taxation of the payout as dividend income for shareholders, particularly if other financing is raised simultaneously to fund investments.

[3] In Bhattacharya (1979), this component is the expected value of the potential costs of financing shortfalls of earnings relative to dividend commitments. These costs are allowed to be arbitrarily small relative to the personal tax cost of dividends, which is unrelated to earnings.

These costly signaling models make predictions about dividend policy that are in line with empirical evidence (Eades, 1982). First, unanticipated increases in dividends should be followed by stock price increases. Second, higher differential personal taxation of dividends, relative to capital gains, should lead to lower dividend payouts because of the increased deadweight cost of the dividend signal. Third, in the cross-section, higher-risk firms should pay lower dividends relative to average earnings because their likelihood of costly cash flow shortfalls is higher. The second comparative static is often *not* true in models in which dividends are paid solely as "residuals" to meet the current consumption needs of investors.[4] Other empirical features of corporate dividend behavior (systematically analyzed in Lintner, 1956), such as their sluggish or discrete adjustments to changes in earnings, are predicted by signaling models in which endogenous investment, chosen following the dividend signal in line with shareholders' interests, plays a major role; see Kumar (1987).

These signaling models require us to modify several conventional notions about the firm's decision criterion and, possibly, the separation of different categories of financing and investment decisions. Foremost among these is the fact that maximization of current market value is no longer an adequate criterion for decision-making at the firm level. The assumed motive for dividend-signaling is the desire of different shareholders to liquidate their holdings at a mix of different times. Thus, assuming a stationary mix of shareholder horizons, some collective weighted average of current and future valuations (and after-tax dividends received) is relevant to decision-making, as Bhattacharya (1979) noted and Miller-Rock (1985) emphasized. Miller and Rock also let the level of investment be determined as a residual by the dividend signal, in contrast to the stance of MM (1961). The resulting signaling cost of underinvestment retains validity in the post-1986 tax regime.

Another salient feature of most costly signaling models of dividends is their assumption that current earnings *cannot* be communicated to outsiders through accounting disclosure. A more reasonable assumption would be that, despite some managerial discretion regarding accounting rules, some coarse indicators of true economic earnings (and, with repetition, their temporal stochastic process) do emerge through the accounting process, and these serve to discipline insiders' incentives to overstate their firm's prospects to raise its value. In Bhattacharya (1980) I considered a "contingent contracting" mechanism in which divergences between dividends or earnings forecasts and realized later earnings lead to revisions of firm values in the equity market. The expectation of this revision process, in turn, leads insiders to pick dividend levels (or earnings forecasts) as signals in line with their "true" earnings prospects, leading to informed valuation.

This model, as noted in the paper, had a major lacuna. Except for special cases (like a random walk in cash flows), there was no guarantee that for all realized future earnings the revised market value would properly reflect the discounted value of further-in-the-future earnings prospects, given the informative self-selection in dividend classes brought about by the value revision prospect. Recent work, by Constantinides and Grundy (1986) among others, has tried to resolve this problem by introducing leveraged securities—such as options, risky debt, and convertibles—as contracting instruments between current shareholders and the market. In the process, a role for capital structure and variable investment choices in the signaling process has emerged.[5] Constantinides and Grundy show that the set of financial contracts that is needed for separating signaling equilibria may be smaller if, out-of-equilibrium, the possibility of non-optimal investment and stock repurchase is admitted.

---

[4] It is easily checked that in the model of DeAngelo (1987), which is a general equilibrium extension of Auerbach (1984), increased personal taxation of dividend income would lead to an increase in before-tax dividends, if consumption in each period is a normal good.

As the above discussion illustrates, we have come a long way in understanding the various tradeoffs associated with debt and dividend policy choices, in an integrated fashion. What is needed now is a synthesis and judgments regarding the range of relevance of different theories, in explaining corporate financial policy and valuation in cross-sectionally and temporally different circumstances (see Scholes and Wolfson, 1988, and Stiglitz, 1985, for attempts). In what follows, we shift our focus to a different set of issues related to the MM (1958) leverage indifference propositions pertaining to the impact of inflation and nominal debt contracts on equity valuation.

Inflation, Leverage, and Stock Market Efficiency

A major empirical puzzle in corporate finance has been the reaction of stock prices and returns to anticipated inflation and unanticipated innovations therein. As Pindyck (1984) notes, the price-level adjusted real decline of the value-weighted New York Stock Exchange index between 1965–81 was nearly 68 percent. Since then, prior to the declines in the summer of 1987 and the dramatic events of October, the index rose to almost three times its value in 1981. These contrasting periods of time have also coincided with sharp (by postwar U.S. standards) accelerations and decelerations in the inflation rate in the economy. Many people would consider such variations in stock prices to be excessive relative to variations in economic fundamentals over these periods, barring major structural changes in the real economy.

Other evidence on the contemporaneous impact of inflation on stock and bond market returns is equally puzzling. Fama (1981) has noted that stock returns are negatively related to measures of both currently anticipated inflation, and differences between the actual inflation rate and its anticipated value. These relationships have been confirmed in a vast number of studies, using measures of anticipated inflation based on moving averages of past rates, or those embedded in nominal Treasury bill rates. In bond markets, the yields or average returns of long-term bonds free of default risk barely kept up with anticipated inflation during periods of high inflation. As a result, the after-tax returns to bondholders who are taxed on their nominal interest income decreased with inflation, which lowered the real values of their nominal credit contracts.

---

[5] See also Ross (1977), Stiglitz (1985), Myers (1984), and Asquith and Mullins (1986), for arguments and evidence regarding the importance of different types of asymmetric information and insiders' incentives for the choice of capital structure and the valuation of equity issues.

Various theories have been advanced to account for the reaction of stock prices in the 1970s to inflation, which should have *no* impact on real stock prices in a world where "money is only a veil" for transaction purposes. Modigliani and Cohn (1979) examined some of the more conventional explanations, and then suggested a very unconventional one based on irrationality on the part of a subset of investors. They identified essentially three conventional explanations. First, the current dollar values of nominal corporate depreciation tax shields are lowered and taxes on nominal capital gains are increased by inflation, in an unindexed tax system based on historical cost and valuation bases. Second, and in contrast, the real after-tax cost of corporate debt is lowered by high inflation, since the full nominal interest is deductible and this seems to rise at most by the amount of anticipated inflation. Third, it is possible that inflationary times are associated with higher riskiness of real activity (like the oil "supply shocks" of the 1970s), which increases the discount rates applied to equity payoffs.

Modigliani and Cohn (1979) have documented that the tax-related first and second effects above largely canceled out each other in the 1970s, producing a stable ratio of corporate taxes as a proportion of total capital valued at replacement cost. Furthermore, using accounting-based measures of operating earnings and their variability as a proportion of replacement value of capital stock, they (and Pindyck, 1984) could not find evidence of sufficient declines in average profitability or increase in the risk thereof to explain the dramatic stock price declines in the 1970s and early 1980s. This motivated the Modigliani-Cohn alternative explanation, based on *misperceptions* of the impact of inflation and corporate leverage on economic earnings on the part of (some) investors and analysts. The essence of their thesis can be summarized through the following example.

Suppose that a firm has perpetual riskless operating income of \$100 per year, of which \$50 is paid as 10 percent interest on \$500 of debt, leaving \$50 in pre-tax earnings and \$25 in after-tax earnings, given the corporate tax rate $T_C = .5$. Assume that, as in a Miller (1977) scenario of zero net tax-shield of debt, the required return on riskless payoffs to equity is 5 percent, because equity earnings are reinvested to generate largely untaxed (and postponable) capital gains. Thus, these after-tax earnings sustain an equity value of \$500 (\$25/.05). Now assume that the (anticipated) inflation rate goes up from 0 to 5 percent, and that the nominal interest rate on the firm's debt goes up to only 15 percent from 10 percent. For convenience, assume that the firm's nominal operating income now increases over time at the rate of inflation, so that in the first period the after-tax nominal accounting profit to equity is .5 (100 − 75) = \$12.5, in the second period .5 (105 − 75) = \$15, in the third period .5 (110.25 − 75) = \$17.625 etc., where \$75 is the new level of annual interest payments in the absence of any additional borrowing by the firm.

What should the value of the firm's equity be now? If investors apply the old *real* required return of 5 percent on equity to the first-period earnings of \$12.5, they get a value of only $250. If the required nominal return on equity is now 10 percent and this is applied to the growing nominal earnings stream starting at $12.5, with a nominal growth rate that is 20 percent in the first period but progressively lower thereafter, we get yet a different answer. The first "answer"—which exemplifies the sort of error Modigliani-Cohn accuse (some) analysts of making—is especially puzzling, since a reduction in the firm's inflation-adjusted after-tax cost of borrowing from .5 (10 percent) - 5 percent to (.5 (15 percent) - 5 percent) = 2.5 percent has had the effect of *lowering* equity value!

An economically correct valuation procedure has to recognize that the firm can refinance 5 percent of its debt each period—because the real value of its debt obligation declines at that rate—and thus keep the ratio between debt interest and nominal operating income constant over time. As a result, the ratio of *net* (of refinancing) payout to creditors before taxes to operating income is also kept equal to its old pre-inflationary level. Indeed, given the lowering of the real after-tax cost of debt relative to the required real return on equity, the firm should increase its leverage even more, up to the point where other deadweight costs of debt cancel its new-found tax advantage for valuation. Even the first step, however, gives the result that equity's economic earning in the first year is (12.5 + 25) = $37.5, growing at a nominal rate of 5 percent, so that applying either a real (5 percent) or nominal (10 percent) required return on equity leads to equity value of $750. This *increase* in rational equity valuation by $250 (750 - 500), reflects the perpetual lowering of the after-tax net payout to creditors by $12.5 (25 - (37.5 - 25)) per period in today's dollars, discounted at 5 percent.

What economic phenomenon can prevent classic MM (1958) homemade investor arbitrage, vis-a-vis an unlevered firm in the same risk class valued at $500, from restoring the above rational valuation of $750? There are two major alternatives. First, if inflation is expected to accelerate, and this worsens the current under-valuation at $250 by 7.5 percent per period in constant dollars, an equity investor would get a net return of only 5 percent (37.5/250 - 7.5) percent on his $250 investment in the equity of the levered firm. Second, if the inflationary scenario is associated with (higher) riskiness of equity returns, then our investor's increased average return of 37.5/250 = 12.5 percent on equity could be required to compensate him for bearing the added risk. The first scenario corresponds, loosely speaking, to that of a "rational speculative bubble." The second scenario might arise from either genuine increases in real asset (cash flow) risk, which would also lower the value of unlevered firms, or risk induced by the presence of "irrational noise traders" whose forecasts regarding *future* stock prices of the levered firm are triggered by a variable such as inflation. We briefly clarify these alternatives before proceeding further.

A rational speculative bubble corresponds to a situation where the price P_t for a dividend stream $\{d_{t+1}, d_{t+2} \ldots\}$, equals the appropriate risk-adjusted discounted value of expected future dividends plus a bubble $\tilde{b}_t$, with the feature: $E(\tilde{b}_{t+1}) = (1 + R)b_t$, where R is the required expected rate of return. In the absence of infinitely-lived utility-maximizing investors, such bubbles can sometimes survive in equilibrium, for example if R is less than the growth-rate of population in an overlapping-generations economy. Notice that the bubble $\tilde{b}_t$ need not be deterministic, and its collapse ($b_t = 0$) may be correlated with other economic variables such as inflation. However, a negative rational bubble ($b_t < 0$) congruent with the Modigliani-Cohn thesis, is inconsistent with market equilibrium since stock prices would then become negative with non-zero probability in finite time, if R is strictly positive and the firm's dividend growth-rate is less than R.

In contrast to a rational speculative bubble, noise trader effects on stock prices may be brought about by a subset of *irrational* agents acting on the basis of whimsical forecasts of future stock prices. The random variability of their resulting trades and the price variability induced thereby may prevent risk-averse rational agents (statistical forecasters) with finite horizons from fully reversing the influence of such trades on prices, as Shiller (1984) pointed out. If the degree of optimism or pessimism on the part of these noise traders is correlated with some uncertain economic variable such as inflation, and this variable inherently has some intertemporal persistence (serial correlation), then the under- or over-valuation produced by their presence may persist over time (Campbell and Kyle, 1986).[6] In the presence of such noise trader effects, the market returns on the equities of a levered and an unlevered firm whose operating cash flows are perfectly correlated, would no longer lie in the same MM (1958) risk class.

Empirically, models using *market* measures of risk have shown some success in explaining the relations between stock prices and anticipated inflation observed in the 1970s. Pindyck (1984) finds that an exponentially weighted moving average of NYSE stock rate of return variances nearly doubled between 1965 and 1981. Based on this, and a portfolio choice model of rational stock market equilibrium with homogeneous investor beliefs, he concludes that the real price decline of 68 percent between 1965 and 1981 could be a combination of around 18 percent owing to average corporate profitability decline, and around 50 percent owing to the doubling of risk, which assumes an empirically plausible relative risk-aversion coefficient for investors of between 3 and 5. Using studies of stock market returns, Chan, Chen, and Hsieh (1985) find that use of a forward-looking proxy measure of riskiness such as the difference between the yields of low-grade (BAA and under) and government bonds or bills, virtually eliminates any impact of *anticipated* inflation on stock returns. In interpreting this result, they suggest that this yield difference is proxying for the default risk-premium on corporate bonds, which increases when firm values are (proportionately) more volatile over time.

A closer look at the evidence, however, raises doubts about both these above rational valuation explanations. Poterba and Summers (1986) find that changes in stock return variances attenuate much more rapidly over time than implied by Pindyck's (1984) model, and hence their highest estimates of the elasticity of stock price response to current volatility changes are around −.225, in contrast to Pindyck's calculation of nearly −1. Campbell (1987) and Shanken (1987) have established that, in response to changes in the Treasury bill rate (which moves closely with expected inflation), the conditional expected returns and variances (or market betas) of stocks move in *opposite* directions. This makes it difficult to rationalize the observed negative correlation between stock returns and the Treasury bill rate (and expected inflation) in terms of associated changes in the riskiness of stock returns. The cross-sectional evidence of Modigliani and Cohn (1980), on valuation decreases in the 1970s across firms differing in their systematic market (beta) risk, also runs counter to the Pindyck explanation, which would have predicted greater value declines for riskier (higher beta) firms.

---

[6] Given sufficiently irrational average optimism about future asset prices, these noise traders may even make higher average returns by taking larger positions in risky assets (DeLong et al., 1987).

Recent evidence, due to DeBondt and Thaler (1987) among others, provides some indirect support for the Shiller (1984) thesis on the impact of noise traders on stock market prices. DeBondt and Thaler find that stock returns display significant negative serial correlation over long horizons, such as 5 years, in that portfolios of past "losers" significantly outperform those of past "winners." These reversals do not seem to be easily explainable by variations in riskiness of these portfolio returns,[7] and they appear to be accompanied by contemporaneous reversals in the rank order of these firms' earnings. These findings led DeBondt and Thaler (1987) to suggest that some parts of the stock market *overreact* to current earnings, and do not recognize the mean reversion implicit in their intertemporal evolution in arriving at valuation.

Further empirical evidence is to me interpretable as being supportive of a story of irrational valuation related to changes in anticipated inflation. Shanken (1987) finds that the positive impact of a higher Treasury bill rate on market return betas or variances of stock returns declines with the market size of the firms in a rather dramatic fashion, whereas the negative impact of the T-bill rate on future expected returns of stocks is almost uniform across firm size categories. Of course, it is precisely smaller firms that are likely to be less well analyzed, or subject to difficulties such as credit rationing (Stiglitz, 1985) in refinancing the inflationary component of their nominal interest payments.[8] Chan, Chen, and Hsieh (1985), using a multifactor model of risk-return tradeoffs in the capital market, are able to largely "explain" the impact of firm size, and expected inflation, on expected returns of stocks. However, they find that an explanatory yield differential (BAA- government) risk factor has a significant January seasonal, as does the raw impact of the firm size variable on stock returns, again suggesting that this variable is not serving solely as a rational measure of default risk on bonds.[9]

---

[7] Bhattacharya (1978) displays a model in which rational valuation with mean-reverting cash flows results in stock return risks (market betas) moving *positively* with stock value, in contrast to the impact of leverage on equity return risk, and in opposition to the risk effect needed to explain the DeBondt-Thaler findings in an efficient market context.

[8] The patterns observed by Shanken are consistent with valuation decreases, produced by increases in interest rates and expected inflation, that are proportionately much greater for smaller firms. In contrast, the negative impact of higher T-bill rates on required future expected return on equities, which is likely to be influenced by substitution away from monetary assets, shows little variation with firm size.

[9] In their study, smaller firms' returns are far more highly correlated with changes in this yield differential risk-factor, especially in January. It is conceivable that greater interest-rate-related valuation changes of small firms (fn. 8) plays a role in this phenomenon.

The accumulation of evidence above presents, in my view, a murky picture vis-à-vis the prevalence of rational (informationally efficient) valuation in the stock market. While earlier time-series empirical studies claiming the presence of excessive volatility (Leroy and Porter, 1981; Shiller, 1981) or inflation-related irrationality (Modigliani and Cohn, 1979) in stock prices were flawed methodologically,[10] more recent evidence on equity valuation is harder to rationalize in terms of the evolution of fundamental technological and preference parameters. As one who as observed the U.S. stock market over the last 15 years, and tried to teach the rational valuation concepts in Modigliani and Cohn (1979) to M.B.A. students, I am inclined to be sympathetic to the likelihood of some truth in their hypothesis of (leverage-related) irrationality in equity under valuation in the inflationary 1970s, and the related rebound over 1982–87.

■ *I thank the editors of this journal for their very helpful comments, and Merton Miller and Franco Modigliani for the education and inspiration they have generously provided over many years.*

---

[10] Criticisms of the excessive volatility tests of Leroy-Porter and Shiller have focused in large part on the inadmissibility of their variance estimators for non-stationary time series (Kleidon, 1986). The regression tests of Modigliani-Cohn, based on the Price-Earnings ratios of NYSE stocks as the dependent variable, have been accused of bias due to omitted variables such as appropriate measures of the riskiness of equity returns. The methodology of Campbell and Shiller (1987) offers some hope of resolving and unifying our understanding of these puzzles.

References

Asquith, Paul, and David Mullins, "Equity Issues and Offering Dilution," *Journal of Financial Economics*, 1986, *15*, 61–89.

Auerbach, Alan J., "Taxes, Firm Financial Policy, and the Cost of Capital," *Journal of Public Economics*, 1984, *16*, 27–57.

Bhattacharya, Sudipto, "Project Valuation with Mean-Reverting Cash Flow Streams," *Journal of Finance*, 1978, *33*, 1317–31.

Bhattacharya, Sudipto, "Imperfect Information, Dividend Policy, and the 'Bird in the Hand' Fallacy," *Bell Journal of Economics*, 1979, *10*, 259–270.

Bhattacharya, Sudipto, "Nondissipative Signaling Structures and Dividend Policy," *Quarterly Journal of Economics*, 1980, *95*, 1–24.

Bhattacharya, Sudipto, "Financial Markets and Incomplete Information: A Review of Some Recent Developments." In Bhattacharya, Sudipto, and George M. Constantinides, eds., *Frontiers of Financial Theory*. Totowa, NJ: Littlefield and Adams, 1988.

Buser, Steven A., and Patrick J. Hess, "Empirical Determinants of the Relative Yields on Taxable and Tax-Exempt Securities," *Journal of Financial Economics*, 1986, *17*, 335–56.

Campbell, John, "Stock Returns and the Term Structure," *Journal of Financial Economics*, 1987, *18*, 373–400.

Campbell, John, and Albert S. Kyle, "Smart Money, Noise Trading, and Stock Price Behavior," Princeton University, mimeo, 1986.

Campbell, John, and Robert J. Shiller, "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," Princeton University, mimeo, 1987.

Chan, K. C., Nai-fu Chen, and David A. Hsieh, "An Exploratory Investigation of the Firm Size Effect," *Journal of Financial Economics*, 1985, *14*, 451–472.

Constantinides, George M., and Bruce Grundy, "Optimal Investment with Stock Repurchase and Financing as Signals," University of Chicago, mimeo, 1986.

DeAngelo, Harry, "Dividend Policy and Personal Taxes," University of Rochester, mimeo, 1987.

DeBondt, Werner F. M., and Richard H. Thaler, "Further Evidence on Investor Overreaction and Stock Market Seasonality," *Journal of Finance*, 1987, *42* 557–81.

DeLong, J. Bradford, Andrei, Shleifer, Lawrence H. Summers, and Robert J. Waldman, "The Economic Consequences of Noise Traders," 1987, CRSP Working Paper 218, University of Chicago.

Eades, Kenneth M. "Empirical Evidence on Dividends as a Signal of Firm Value," *Journal of Financial and Quantitative Analysis*, 1982, *17*, 471–500.

Fama, Eugene F., "Stock Returns, Real Activity, and Inflation," *American Economic Review*, 1981, *71*, 545–65.

John, Kose, and Joseph T. Williams, "Dividends, Dilution, and Taxes: A Signaling Equilibrium," *Journal of Finance*, 1985, *40*, 1053–70.

Keim, Donald B., "Dividend Yields and Stock Returns: Implications of Abnormal January Returns," *Journal of Financial Economics*, 1985, *14*, 473–90.

Kim, E. Han, "Optimal Capital Structure in Miller's Equilibrium." In Bhattacharya, Sudipto, and George M. Constantinides, eds., *Frontiers of Financial Theory*. Totowa, NJ: Littlefield and Adams, 1988.

Kleidon, Allan, "Anomalies in Financial Economics: Blueprint for Change?" *Journal of Business*, 1986, *59*, S469–S499.

Kumar, Praveen, "Shareholder-Manager Conflict and the Information Content of Dividends," 1987, Carnegie-Mellon University, mimeo.

Leroy, Stephen F. and Richard D. Porter, "The Present-Value Relation: Test Based on Implied Variance Bounds," *Econometrica*, 1981, *49*, 555–74.

Linter, John, "Distributions of Incomes of Corporations Among Dividends, Retained Earnings, and Taxes," *American Economic Review*, 1956, *46*, 97–113.

Litzenberger, Robert H., and Krishna Ramaswamy, "Dividends, Short Selling Restrictions, Tax-Induced Investor Clienteles, and Market Equilibrium," *Journal of Finance*, 1980, *35*, 467–482.

Miller, Merton H., "Debt and Taxes," *Journal of Finance*, 1977, *32*, 261–275.

Miller, Merton H., "The Informational Content of Dividends." In Dornbusch, Rudiger, Stanley Fischer, and John Bossons, eds., *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*. Cambridge, MA: MIT Press, 1987, pp. 37–58.

Miller, Merton H., and Franco Modigliani, "Dividend Policy, Growth, and the Valuation of Shares," *Journal of Business*, 1961, *34*, 411–433.

Miller, Merton H., and Franco Modigliani, "Some Estimates of the Cost of Capital to the Utility Industry, 1954-7," *American Economic Review*, 1966, *56*, 333–391.

Miller, Meton H., and Kevin Rock, "Dividend Policy under Asymmetric information," *Journal of Finance*, 1985, *40*, 1031–1051.

Miller, Meton H., and Myron Scholes, "Dividends and Taxes," *Journal of Financial Economics*, 1978, *6*, 333–364.

Miller, Merton H., and Myron Scholes, "Dividends and Taxes: Some Empirical Evidence," *Journal of Political Economy*, 1982, *90*, 1118–1141.

Modigliani, Franco, "Debt, Dividend Policy, Taxes, Inflation, and Market Valuation," *Journal of Finance*, 1982, *37*, 255–273.

Modigliani, Franco, and Richard Cohn, "Inflation, Rational Valuation, and the Market," *Financial Analysts Journal*, March/April 1979, *35*, 24–44.

Modigliani, Franco, and Richard Cohn, "Inflation and the Stock Market." In Boeckh, J. Anthony, and Richard T. Coghlan, eds., *The Stock Market and Inflation*. Homewood, IL: Dow Jones-Irwin, 1980, 99–117.

Modigliani, Franco, and Merton H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review*, 1958, *48*, 267–297.

Modigliani, Franco, and Merton H. Miller, "Corporate Income Taxes and the Cost of Capital: A Correction," *American Economic Review*, 1963, *53*, 433–443.

Myers, Stewart C., "The Capital Structure Puzzle," *Journal of Finance*, 1984, *39*, 575–592,

Pindyck, Robert S., "Risk, Inflation, and the Stock Market," *American Economic Review*, 1984, *74*, 335–351.

Poterba, James, and Laurence Summers, "The Persistence of Volatility and Stock Market Fluctuations," *American Economic Review*, 1986, *76*, 1142–1151.

Ross, Stephen A., "The Determination of Financial Structure: The Incentive-Signalling Approach," *Bell Journal of Economics*, 1977, *8*, 23–40.

Scholes, Myron, and Mark Wolfson, "Issues in the Theory of Optimal Capital Structure." In Bhattacharya, Sudipto, and George M. Constantinides, eds., *Frontiers of Financial Theory*. Totowa, NJ: Littlefield and Adams, 1988.

Shanken, Jay, "The Intertemporal Capital Asset Pricing Model: An Empirical Investigation," 1987, University of Rochester, mimeo.

Shiller, Robert J., "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?" *American Economic Review*, 1981, *7*, 421–436.

Shiller, Robert J., "Stock Prices and Social Dynamics," *Brookings Papers on Economic Activity*, 1984, *2*, 457–498.

Spence, A. Michael, "Job Market Signalling," *Quarterly Journal of Economics*, 1973, *87*, 355–379.

Stiglitz, Joseph E., "Taxation, Corporate Financial Policy, and the Cost of Capital," *Journal of Public Economics*, 1973, *2*, 1–34.

Stiglitz, Joseph E., "Credit Markets and the Control of Capital." *Journal of Money, Credit, and Banking*, 1985, *17*, 133–52.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 149–158

MM—Past, Present, Future

Franco Modigliani

Before getting serious about the MM theorem's past and future, I would like to say a few words about its more intimate aspects—MM *chez-soi* as it were.

Past—Some Reminiscences

One of the most productive days in my life was the day in 1956 when it was agreed that Merton Miller—then a shy and retiring young man (would you believe it!)—would sit in my graduate course in "Money and Macroeconomics" at Carnegie Institute of Technology. It was an exciting course with excellent students. It took a broad view of the subject. One issue that was covered was the cost of capital as a determinant of the rate of investment. I had been intrigued by the subject ever since attending a National Bureau conference on Business Finance at which I gave a (fairly conventional) paper (Modigliani and Ziman, 1952). But mostly I listened to a paper by David Durand (1952) in which the possibility that financial structure would not affect the market valuation or the cost of capital was suggested, only to be rejected as not relevant to the actual capital markets.

In preparing my lecture dealing with the cost of capital, I was able to provide (for a world of no taxes) a sort of proof of Proposition I, based on arbitrage. I reported the result to my class the next day, adding that I didn't really believe my result and there probably was something wrong. But Miller was instantly captured by the result because, he said, he knew of a recent paper (Allen, 1954) which provided empirical support for the result, even though it lacked a convincing rationale. We thus formed,

■ *Franco Modigliani is Institute Professor and Professor of Economics and Finance, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts.*

there and then, the MM alliance, pledged to go to the bottom of the issue. Except for the fortuitous presence of Miller in this class and his immediate appreciation for the importance of this issue, the MM proposition (Modigliani and Miller, 1958) might never have seen the light. As explained by Miller, the second paper on dividend policy (Miller and Modigliani, 1961) was a derivative of the first, initially spurred by the intention of completing and refining our proof.

The MM paper is unquestionably the most popular of my writings, primarily because it has been, and continues to be, required reading for many graduate business schools. The way it is used in master's programs, as far as I can ascertain, is as a whipping boy—the instructor assigns the paper for reading and then has a feast tearing it to shreds. Wherever I have gone I have found former MBA students recognizing my name and quickly relating it to their MM days. But I have discovered that MM is particularly popular among Washington's young taxi drivers, since an extraordinarily large portion of them seem to be African students going to graduate business schools. Whenever I take a taxi in Washington and find that it is driven by a young man with an accent, I try to bring the subject around to MM and the almost universal result is instant kinship, together with a complaint—one that is actually fairly universal among former MBA students—that the article is much too difficult, though I suspect that they enjoy talking about the article just because it was so difficult (cognitive dissonance!). I am told its essence could have been restated in a more comprehensible fashion.

I have always been ready to apologize for causing trouble, and have offered explanations in our defense. Surely now that the MM theorem seems almost trivial, one could easily present a proof in a more readily intelligible fashion. But not then, when colleagues (academic and practitioner) took it as self-evident that there was a unique, value-maximizing debt ratio and regarded our propositions as plainly preposterous. There was therefore a need for rigorous and varied arguments to show the formal proof of the result and provide the "new" intuition as to why it made sense. Thus the article had never been meant as an expository paper for master's students in business. It was addressed to finance specialists and it was written tongue-in-cheek, not really to demonstrate that leverage could not possibly affect market values in the actual world but to shock those who accepted the then-current naive view that some debt in the capital structure had to reduce the cost of capital even in the absence of taxes simply because the interest rate was lower than the earnings-price ratio on equity.

Present and Future

Miller has reviewed, in his usual masterly fashion, the validity and working of the MM theorem under conditions of perfect market and no taxes. There seems little to add to that subject, especially since, by now, countless alternative proofs of the theorem have been provided.

So I would like to come back to some issues that are still with us—and probably will be for a long time. How should the MM propositions be modified in the presence

In MMC we rejected this result on the ground that the stream of annual tax saving $\tau_c rD$, in contrast to the profit stream $(1 - \tau_c)\bar{X}Z$ was "a sure stream" (MMC, p. 435); therefore it should be capitalized not at ρ but at the sure rate r. In that case, the last term of (3) becomes $(\tau_c rD)/r = \tau_c D$, (cf. MMC equation 3) which is the basic contribution of MMC. With the corporate rate nearly 50 percent at the time of our writing, a dollar of debt would raise the market value of the firm by roughly 50 cents. But this result rests on the assumption that the tax saving stream $\tau_c rD$ is constant, perpetual, and absolutely certain like the coupon of a government bond. In MMC we did mention some limitations to the validity of the assumption, on account of the possibility of changes in the tax code as well as of profits falling below contractual interest. But the much more fundamental objection relates to the assumption that leverage policy can be modeled as the choice of an amount of debt in the capital structure, fixed once and for all. This assumption seems untenable in a world in which the movement of expected profit and size of the firm is widely supposed to follow something like a random walk (or a martingale). It seems much more reasonable to suppose that the leverage policy of the representative firm can be described as aiming at maintaining the debt in a stable relation to the scale of the firm as seen at any given date. If, for example, the appropriate scale variable is approximated by the net of tax cash flow of the (unlevered) firm (or equivalently by market value of the firm V_u), we can write

$$\tilde{D} = d(1 - \tau_c)\bar{X}Z \tag{4}$$

where d embodies the firm's debt policy, and the superscript ~ on D is to remind us that future debt, like future profit, is a random variable. Under these conditions total after-tax return of the levered firm X^τ, given by (1), can be restated as:

$$X^\tau = (1 - \tau_c)\bar{X}Z + \tau_c rd(1 - \tau_c)\bar{X}Z = (1 - \tau_c)\bar{X}Z(1 + \tau_c rd). \tag{5}$$

The essential difference with equation (1) above is that according to (5), the return of the levered firm can be seen as proportional to that of the unlevered firm, $(1 - \tau_c)\bar{X}Z$, the proportionality factor being $(1 + \tau_c rd)$. This proportionality of returns in turn implies that the market value of the levered firm V_L must also be proportional to that of the unlevered firm, V_u, given by equation (2), with the same proportionality factor. Thus

$$\begin{aligned} V_L &= V_u(1 + \tau_c rd) = V_u + \tau_c rdV_u \\ &= V_u + \tau_c rd\left[(1 - \tau_c)\bar{X}Z/\rho\right]; \end{aligned} \tag{6}$$

or using (4),

$$V_L = V_u + \tau_c rD/\rho \tag{7}$$

According to (7) the value of tax saving comes to $(r/\rho)\tau_c D$ which is much smaller than MMC's $\tau_c D$. In fact, for the postwar period, it is only about 1/3 as large.

One particularly intriguing implication of (7) is that the value of leverage it implies is the same as that implied by the original MM (see equation (3) above). In other words, if one accepts the reasonable notion that the appropriate discount rate for $\tau_c rD$ is ρ rather than r, then the correction paper and its "definitive" corrections need never have been written: Personal taxation aside, the definitive truth was all in MM (though the original way of establishing the result was defective). Of course this is somewhat of an exaggeration since it would be foolhardy to claim that ρ is the appropriate way to discount the tax saving under all circumstances and tax regimes (for example, independently of loss carryover and loss salability provisions). In particular, a number of considerations suggest that the tax saving might be even riskier than the basic cash flow, because of possible changes in interest rates, leverage policy, taxation, and so on. Perhaps if we had been very wise we should have presented the basic valuation formula in a form more general than either MMC or the proposed alternative given by (7) (and the original MM), namely as

$$V_L = V_u + \chi \tau_c rD \tag{8}$$

Here χ denotes the reciprocal of the discount rate to be applied to $\tau_c rD$, which could range from r to ρ or even higher. Note that (8) retains the MM assumption that the gain from leverage is proportional to D. This may not be strictly true when we take account of bankruptcy cost, and the probability that taxable profits (as reduced by investment tax credit and other factors) may fall short of interest. As a result, V_L might increase with D but at a decreasing rate, at least beyond some point.

But even if MM was more nearly correct than MMC in the treatment of tax saving, it did contain a second error as suggested earlier—namely, neglecting the role of differential personal taxation of the different sources of return produced by a corporation. Clearly, the value of leverage to investors must depend on the amount of additional taxes, corporate and personal, that could be avoided by paying out an additional dollar of corporate returns in the form of interest (which is subject to the personal income tax) while reducing by a dollar stockholders profits (which are subject to the corporate plus the appropriate personal income taxes).

There is presumably broad agreement ever since the Farrar, Selwyn contribution (1967) that under the U.S. system of taxation, what leverage adds to the investor's stream of returns net of both corporate and personal taxes can be written as the second term of (8) but with the corporate tax rate, τ_c, replaced by a different coefficient, say

$$l = 1 - (1 - \tau_c)(1 - \tau_g)/(1 - \tau_p) \tag{9}$$

where τ_p and τ_g are respectively the personal income and capital gain tax rate. This expression coincides with the coefficient of BL in Miller's equation for the gains from leverage in this volume (except for notational differences).

The common sense of (9) is that if a dollar of profit is received as return on equity, and assuming that none is paid out in the form of dividends, then the tax that will be paid is first the corporate tax τ_c and then the capital gains tax τ_g leaving $(1 - \tau_c)(1 - \tau_g)$. On the other hand, a dollar of interest will be taxed only at the personal level, leaving $(1 - \tau_p)$. In the limiting case when there are no personal taxes or, more generally, capital gains are taxed like all other personal income, that is, $\tau_g = \tau_p$, it can be seen that l reduces to MM's τ_c because it makes no difference in what form corporate earnings are paid out—leverage is valuable only because it saves corporate income taxes. Thus we are back to the MM results. However, if $\tau_g < \tau_p$, then $l < \tau_c$ and could even be zero if, for example, τ_g were zero and $\tau_p = \tau_c$.

However, despite the agreement about l, there is room for disagreement as to what it implies for the market valuation of levered stock. Miller (and others) interpret his above-mentioned equation as indicating what a particular investor would stand to gain or lose from corporate versus personal leverage, given his specific marginal tax rates. He then relies on the notion that the equilibrium market valuation must be determined by the marginal holder and on the existence of tax-exempt securities to conclude that the marginal holders, who determine the market value of l, may (and will tend to) be characterized by a set of individual marginal tax parameters such that $\tau_p = \tau_c$. Combined with the assumption that τ_g is small, this condition, as we have just shown, implies that the market value of leverage, l, is close to zero. In equilibrium leverage is, taxwise, basically worthless.

In my own analysis, inspired by a number of earlier contributions (Brennan, 1970; Elton and Gruber, 1974; Auerbach and King, 1982) I have relied on the mean-variance framework and shown that, because of the benefits of portfolio diversification, it implies that every investor will, in principle, be led to hold a position (possibly negative) in a wide variety of stocks (in principle all), albeit the stocks will be held in different combinations reflecting tax parameters as well as risk preferences. Under these conditions, it turns out that in (8) above, τ_c is still replaced by l, but l is evaluated using not the parameters of the marginal holders as in Miller, but rather an average of all relevant marginal tax rates (since basically everyone is holding all the stocks) but weighted by both tax parameters and risk aversion, which contribute to determine individual holdings.

Since up to the latest tax reform the average capital gain tax was much smaller than the average personal tax rate, my approach implies that l is substantially positive, though well below τ_c. In MO it was estimated that, for the postwar years up to 1988, the coefficient l—computed from estimated average tax rates—could be placed around 0.33. Accordingly, the value of tax saving could be estimated at 0.33 if the saving were capitalized at the sure rate, but only at 0.11 if capitalized at the more relevant risky rate. This last estimate implies that a 10 percent rise in leverage would increase market valuation by just over one percent (instead of some 4 percent if tax savings are capitalized at the sure rate). Obviously this effect is small enough to suggest that, up to the Tax Reform Act (TRA), leverage policy decisions must have been largely swayed by considerations other than tax saving. Note that this conclusion is not very different from that reached by Miller in MI, but it is based on an entirely different reasoning which has quite different implications with respect to the effect of changes in tax laws, such as the most recent one.

One salient nature of this revision is that all sources of income from the corporation—capital gains, interest, dividends—are taxed basically in the same way. Only the capital gains tax remains marginally lower because of the possibility of postponing realization (and the postponement may be quite long under the so-called death loophole). Therefore, to a first approximation, we are back in the regime implicitly assumed in both MM and MMC for which l is approximately τ_c. Note that this conclusion holds whether one accepts Miller's formulation relying on marginal taxes or ours, based on average tax rate, since $l = \tau_c$ for every individual (including tax exempt institutions), and τ_c is of course the same for all, namely 0.34, under the new law.

Miller seems to view this feature of new tax law with alarm for it changes his value of l very significantly: from zero, or close to, in the old regime, to τ_c or 0.34, under the new tax system. He thinks that corporations might engage in an extensive search for types of organizations other than the standard corporate form in response to the loss of favorable tax treatment of capital gains due to retained earnings. He further suggests that they might respond to the opportunity to reduce taxes through leverage, by "gut[ting] the corporate tax with high leverage capital structure." He even suggests that the Tax Reform Act may provide a unique opportunity to test these expectations.

However, according to my analysis, the effects of TRA on leverage should be negligible. The reason is that though Miller and I agree that under the new regime $l = \tau_c = 0.34$, it happens that in the old regime, my estimate of l is very nearly the same number, namely, 1/3 (see above). This is the result of the fact that, although τ_c was close to 0.5, my estimate of l, based on average tax rates, was appreciably lower than l. If the tax gains are capitalized at the risky rate, then both the above figures should be divided by 3. Clearly, there is essentially no change in the value of leverage on account of TRA, and therefore, according to my analysis, no reason to expect any changes in leverage policy. There continues to be an advantage to leverage, though rather small.

With such sharp differences in implications, perhaps the next few years may serve as a crucial experiment. But I don't really expect that we will be so lucky as to witness the experiment. No doubt there will be numerous interferences from exogenous shocks, such as further changes in tax laws, or changes in incentives to leverage other than income taxes, not to mention the possibility of inconclusive results.

The Tax Effects of Dividend Policy

We can readily establish that, at least under the MM (1961) assumption that investment policy is independent of dividend policy, a company paying Δ dollars of dividends will increase the flow of taxes and reduce the net of tax stream received by the public by $(\tau_p - \tau_g)\Delta$—at least as long as the alternative to paying dividends is to buy back equity. The effect on the value of a corporation can then be assessed by capitalizing the stream of tax losses. If we are prepared to assume that the current dividend Δ will be paid forever and with complete certainty, then it would be appropriate to capitalize the tax loss at the sure rate r, following the procedure used by MM in capitalizing the stream of tax saving $\tau_c rD$. The contribution to value from dividends would then come to $(\tau_g - \tau_p)\Delta/r$, implying a very large negative effect—something like a reduction in market value of $8.00 per dollar of dividend, or a decline in value of the order of 8 percent for a 10 percent increase in payout ratio.

But again, there seems to be absolutely no justification for the assumption that incremental tax dollars due to the current dividend will continue forever and with as much certainty as if they represented the interest on a government bond—barring limiting cases where the income itself is certain. How can the dividends be more certain than the stochastic income out of which they are generated? This suggests as one possible reasonable alternative that what the market should value is dividend policy, defined, say, as the long run pay-out ratio or as the proportion δ of (long run) tax corrected expected profits to be paid out in dividends. (It may be argued that many companies have a short-run policy of stabilizing dividends rather than the payout rate. But this is because payout policies typically call for dividends proportional to smoothed income which is, clearly, not inconsistent with our hypothesis of long-run proportionality of dividends and income.) Under the above hypothesis, the expected flow of dividend induced by the policy becomes a stochastic variable:

$$\tilde{\Delta} = \delta\bar{X}(1 - \tau_c)Z \tag{10}$$

which presumably should be capitalized at the risky rate ρ just like the tax saving from leverage. Accordingly, the loss in value becomes $[(\tau_p - \tau_g)\delta\bar{X}(1 - \tau_c)]/\rho$. This implies that a 10 percent change in the payout δ would reduce market value by less than 3 percent. Needless to say, using ρ as the discount rate seems plausible (a lot more than using r), but alternative formulations could no doubt be defended. Thus, the risk in the tax stream arising from dividends might be greater than for profits because of possible changes in dividend policies (which our analysis suggests should therefore be avoided) or in the tax treatment of capital gains versus ordinary income, of which we have just had a conspicuous example.

If our formulation is reasonably close to truth, then it would suggest that the impact of dividend policy on value by way of tax effect could again easily be swamped by other factors, such as the lack of investment opportunities plus restraint in buying back one's own share because of IRS disapproval or convention, or perhaps even signalling (despite some skepticism on my part).

These considerations bring us back once more to the implication of the grand experiment represented by the TRA. Under the new law, $\tau_p - \tau_g$ will be zero except for the postponability of the tax on capital gains. Thus, the tax cost of paying dividends will have been whittled down to a negligible level, even if not actually to zero. Anyone taking the view that dividends, though quite costly under the old tax law, were nonetheless being extensively paid out (on the average not far from 50 percent of profits) must conclude that with the great reduction in the tax cost of dividends, brought about by the new tax law, there must be a substantial increase in dividend payout. In addition, one should expect other observables, like a rise in the price of firms which were paying large dividends in relation to low-paying firms.

On the other hand, if I am right in holding that the cost of paying dividends was reasonably small under the old tax, then the effect of the new regime should be small at best. For the same reason there should be no significant realignment in the value of firms with different payouts. Here, too, in principle the TRA should provide an opportunity for a crucial experiment, though again I have doubts that the opportunity will be realized. The tax regime may not stay put long enough, and even if it does it may not be easy to interpret observed behavior of boards of directors and markets as unequivocally supporting one model or the other. For instance, little change, which in principle should be interpreted as support of my model, could simply reflect noise in the data, unobservable expectations, or even plain inertia.

And I must recognize in this connection that, ever since my work in analyzing the effects of inflation on market valuation (Modigliani and Cohn, 1979), I have become a bit disenchanted with the indiscriminate use of superrationality as the foundation for models of financial behavior. In particular, I have come to the conclusion that the effects of inflation on valuation derive primarily from the market's failure to undersand how to value equities in the presence of significant inflation, which results in systematic, predictable error. I have supported this contention with evidence that this hypothesis can account for the depressed market of the late 1970s and early 1980s, the recovery until 1986, the bubble of 1986–87, and its recent unavoidable bursting (Modigliani 1979, 1980).

I must finally acknowledge that there is one further reason for little short-run changes in dividends (or even leverage policy) which is consistent with alternative valuation models and rational behavior, namely a widely held belief by the public that equal taxation of capital gains and other personal income will just not last. In an era which, in a matter of a very few years, has witnessed the most remarkable sequence of "permanent" changes in business and personal income taxation, there is certainly ample rational grounds for not responding to the changes implied bv the latest tax bill as though they were as perennial as the Rock of Gibraltar!

■ *I wish to express my thanks to Robert Merton, Enrico Perrotti and Robert Taggart for reading an earlier version of this manuscript and making many valuable suggestions.*

References

Allen, F. B. "Does Going into Debt Lower the 'Cost of Capital'?," *Analysts Journal*, August 1954, *10*, August, 57–61.

Auerbach, Alan J., and Mervyn A. King, "Taxation, Portfolio Choice and Debt-Equity Ratios: A General Equilibrium Model," 1983, *Quarterly Journal of Economics*, *98*, 4, 587–609.

Brennan, M. J., "Taxes, Market Valuation and Corporate Financial Policy," *National Tax Journal*, 1970, *23*, 4, 417–427.

Durand, David, "Cost of Debt and Equity Funds for Business: Trends and Problems of Measurement." In *Conference on Research in Business Finance*, New York: National Bureau of Economic Research, 1952, pp. 215–147.

Elton, Edwin J., and M. Gruber, "Marginal

Stockholder Tax Rates and the Clientele Effect," *Review of Economics and Statistics*, 1970, *52*, 68–74.

Farrar, Donald E., and Lee L. Selwyn, "Taxes, Corporate Financial Policy, and Returns to Investors," *National Tax Journal*, December 1967, *20*, 4, 444–54.

Miller, Merton H., "Debt and Taxes," *The Journal of Finance*, May 1977, *32*, 2, 261–75.

Modigliani, Franco, "Financial Markets," lecture presented for the Industrial Liaison Program, Massachusetts Institute of Technology, Cambridge, MA, 1980, unpublished.

Modigliani, Franco, "Debt, Dividend Policy Taxes, Inflation and Market Valuation," *The Journal Of Finance*, May 1982, *XXXVII*, 2, 255–73.

Modigliani, Franco, and Richard A. Cohn, "Inflation and the Stock Market." In Boeckh, Anthony, and Richard T. Coghlan, eds., *The Stock Market and Inflation*. Homewood, IL: Dow Jones-Irwin 1979, pp. 3–23.

Modigliani, Franco, and Richard A. Cohn, "Inflation, Rational Expectation, and the Market," *Financial Analyst Journal*, March/April 1979, *35*, 2, 24–44.

Modigliani, Franco, and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *The American Economic Review*, June 1958, *XLVIII*, 3, 261–97.

Modigliani, Franco, and Merton H. Miller, "Dividend Policy, Growth, and the Valuation of Shares," *Journal Of Business*, October 1961, *34*, 411–33.

Modigliani, Franco, and Merton H. Miller, "Corporate Income Taxes and the Cost of Capital: A Correction," *American Economic Review*, June 1963, *53*, 3, 433–43.

Modigliani, Franco, and M. Ziman, "The Effect of Availability of Funds and the Terms Thereof, on Business Investment." In *Conference on Research in Business Finance*. New York: National Bureau of Economic Research, 1952.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 159–164

Theory versus Empiricism in Academic Economics: Update and Comparisons

Theodore Morgan

Six years ago last summer, Wassily Leontief, a past president of the American Economic Association and Nobel Laureate, analyzed the kinds of articles published in the *American Economic Review*. He focussed on the proportion of articles devoted to theory as compared to empirical analysis. His findings illustrated his disapproval of the state of academic economics. Leontief (1982, pp. 104–107) wrote:

> Not having been subjected from the outset to the harsh discipline of systematic fact finding, traditionally imposed on and accepted by their colleagues in the natural and historical sciences, economists developed a nearly irresistible predilection for deductive reasoning... The aversion of the great majority of the present-day academic economists for systematic empirical inquiry [is best revealed by] the methodological devices they employ to avoid or cut short the use of concrete factual information. The primary information, however detailed, is packaged in a relatively small number of bundles labeled "capital," "labor," "raw materials," "intermediate goods," "general price level," and so on... Masses of concrete detailed information contained in technical journals, reports of engineering firms, and private marketing organizations, are neglected.

Leontief (p. 107) anticipated that as the profession "lags intellectually," research people in adjoining fields (demography, sociology, political science; ecology, biology, health sciences, engineering, other applied physical sciences) will eventually be vocal in their dismay at the "stable, stationary equilibrium and the splendid isolation" of academic economics.

■ *Theodore Morgan is Professor of Economics, Emeritus, University of Wisconsin, Madison, Wisconsin.*

Table 1
Articles published in the *American Economic Review*

	Leontief data		*new data*
	Mar. 1972 to Dec. '76	*Mar. 1977 to Dec. '81*	*Mar. 1982 to Dec. '86*
Theory—models, analysis, methodology	%	%	%
Mathematical models without any data	50	54	42
Analysis without any mathematical formulation and without data	21	12	7
Statistical methodology	1	1	1
Subtotal	72	66	50
Empirical analysis			
Based on data generated by author's initiative	1	1	5
Based on data published or generated elsewhere	21	23	39
Not using statistical inference	5	7	1
Based on artificial simulations and experiments	1	2	6
Subtotal	28	34	51
Total	100.0	100.0	100.0

The 13 or 14 "articles" in each issue were tabulated, not the "shorter papers." There were Procrustean problems: I tried hard to put each article into one class; but in rare cases, I failed and they were divided between two classes, and in a few cases three. If Presidential addresses, pretty much fixed by the occasion, were omitted, the second subclass would be 1.5 percent smaller. Sometimes I wanted another subclass, "Theories, plus a test of them," but stayed with the Leontief divisions. Two small Leontief divisions are combined into the next-to-last subclass above. All data are rounded to the nearest percent.

Sources: For 1972–76 and 1977–81, Leontief (1982, p. 107). See also *The Economist*, July 17, 1982, p. 67. For 1982–86, my own count, details on request.

In Table 1, I reproduce Leontief's tabulation of *American Economic Review* articles for 1972–76 and 1977–81, and then add new data covering the most recent five years, 1982–86. The question is: what changes are occurring in the proportions of theoretical articles and empirical articles?

From the critics' point of view, there has been improvement. Theory, with its subdivisions, has declined from 72 percent to 66 percent, then to 50 percent. "Mathematical models without data," the form of theory that the critics might grade lowest, rose to a height of 54 percent in the middle five years, and then fell to 42 percent.

Conversely, empirical analysis rose from 28 percent to 34 percent to 51 percent. But this generally "good" news has its soft spots. Most of the increase in empirical papers comes from more papers that use "data published or generated elsewhere."

Table 2

Articles published in the *Economic Journal*

	Mar. 1972 to Dec. 76	*Mar. 1977 to Dec. 81*	*Mar. 1982 to Dec. 86*
Theory—models, analysis, methodology	%	%	%
Mathematical models without any data	34	50	52
Analysis without any mathematical formulation and without data	24	8	7
Statistical methodology	1	0	0
Subtotal	59	57	58
Empirical analysis			
Based on data generated by author's initiative	1	3	2
Based on data published or generated elsewhere	39	37	38
Not using statistical inference	0	1	1
Based on artificial simulations and experiments	1	1	2
Subtotal	41	43	42
Total	100	100	100

As with the *AER* tabulation, only "articles"—fewer in the *EJ*, 4 to 14 in each issue —are counted. "Notes and memoranda," etc., are omitted. There were occasional Procrustean problems, solved as with the *AER*. All percents are rounded.

Leontief complained that the government statistics mainly relied on in these papers were compiled for administrative or business use, not for scientific purposes. The Leontief-style critics would like to see more papers in the star class, "Empirical analysis based on data generated by the author's initiative." That class has risen from (less than) 1 percent in the first ten years to 5 percent in the most recent data; the critics would mutter that is still only one in twenty.

Abroad, The *Economic Journal* in Britain presents the most notable standard available to lay against the levels and trends of the *AER*. The *EJ* is very hospitable to economists in other countries (one article came from a foreign prison); still the decisions are made by its own board of editors. The data of Table 2 cover the same five-year periods.

The *EJ* 's emphasis on theory, all kinds, has stayed remarkably constant at 57 percent to 59 percent. Empirical analysis is equally constant at 41 percent to 43 percent.

Within the theory group there is dramatic change. "Mathematical models without any data" have risen from 34 percent to 52 percent, while "analysis without mathematics or data" fell from 24 percent to 7 percent. Nearly all the change occurred between the first and second five-year periods. Omitting presidential addresses lowers the "analysis, etc." class by 3 percent.

The *EJ* during the past five years has published much more math-without-data, about 23 percent more, than has the *AER*.

Four Other Social and Physical Sciences

Table 3 compares academic economics, from the above data for 1982–86, with two other U.S. social sciences and two U.S. physical sciences. The articles tabulated for the social and physical sciences are those appearing in the official journals of their associations.

For chemistry and physics it was appropriate to group together the first two subclasses under "empirical analysis." These subclasses had a very large share of papers with mathematical, graphic, or tabular constructions closely tied to data.

Academic economics is unique among these five disciplines in the share of attention given to "mathematical models without any data," with the *EJ*'s 52 percent surpassing even the *AER*'s 42 percent. The nearest competitors lag far behind: in political science 18 percent of the papers are mathematical models without data; in physics 12 percent. Sociology and chemistry are at or near zero.

Political science and sociology do have more of "analysis without any mathematical formulation and without data" than does economics (23 percent and 18 percent versus 7 percent, respectively). Chemistry and physics have no papers in this class at all.

In the totals overall for theory papers without evidence, economics remains far ahead (50 percent or 58 percent); with political science behind at 42 percent and the other fields not even in the running.

As to the empirical papers: chemistry is by far the most empirical of these disciplines (rounding gives it 100 percent of all papers!), nearly always relying on the author's experiments and simulations or on the author's other work. Political science and sociology share with economics the pattern of often relying on "data published or generated elsewhere." But those disciplines have two to twelve times the proportion of papers "based on data generated by the author's initiative." (In the numbers of main papers published, there is a sharp difference between the journals of the three social sciences—with each recently between 36 and 55 a year; and those of the two physical sciences—increasing by 1986 to 1100 papers a year (chemistry), or around 13,400 (physics). There are various possible explanations. The contrast remains remarkable.)

In summary: Among these five disciplines, economics and chemistry are extremes. Economics has half or more of papers devoted to theory-without-data; chemistry has none. Economics devotes half or less of its papers to empirical analysis; all the chemistry papers are empirical. The other disciplines string out along the way between these extremes: in order from economics come political science, sociology, and then physics.

Table 3

Main papers published in the official journals of five academic disciplines, 1982–86

	Economics					
	AER	*EJ.*	*Political Sci.*	*Sociology*	*Chemistry*	*Physics*
Theory—models, analysis, methodology	%		%	%	%	%
Mathematical models without any data	42	52	18	1	0	12
Analysis without any mathematical formulation and without data	7	7	23	18	0	0
Statistical methodology	1	0	0	3	0	0
Subtotal	50	58	42	22	0	12
Empirical analysis						
Based on data generated by author's initiative	5	2	10	23	17	41
Based on data published or generated elsewhere	39	38	41	51		
Not using statistical inference	1	1	2	1	0	0
Based on experiments and simulations	6	2	6	3	83	48
Subtotal	51	42	58	78	100	88
Total	100	100	100	100	100	100

The economics data are from Tables 1 and 2. Political science data are from the *American Political Science Review*, published by the American Political Science Association. It is a quarterly, printing 226 articles 1982 through 1986. "Communications," book reviews, and "Controversy" are excluded from the count. Sociology data are from the *American Sociological Review*, published in six issues a year, with 255 articles 1982 through 1986. The *Review* is the official journal of the American Sociological Association. "Comments" and "Research Notes" are excluded from tabulation. The *Journal of the American Chemical Society* is published in 26 issues each year, without separate series for sub-fields. There were about 4920 main papers printed in the five years 1982–86 (estimated from the average of the papers in 1982 and 1986). The papers classified were a random sample: 9% in 1985, 5% in 1986, and 2% in each of the earlier three years—a total of 239 papers. The *Physical Review* is published by the American Physical Society, in four series: General Physics, Condensed Matter, Nuclear Physics, and Particles and Fields. The tabulation above is from a stratified random sample. About 58,740 papers (shorter contributions excluded) were published 1982 through 1986 (estimate based on the average for 1982 and 1986). In 1986 80 papers were sampled (20 in each of the four series); and in 1982–85, 20 were randomly chosen in each year, with a different one of the four fields chosen in each of the four years. Hence a total of 160 papers were classified. All data are rounded to the nearest whole percent.

Conclusions

These are the data. What can we make of them?

It is helpful to try to stand back, try for perspective, much as one might muse over the kind of lectures given and papers written at the University of Paris in 1400. In each discipline, is the most socially useful kind of product being created—I take for granted this is the right objective, as Leontief did also—in view of the problems extant and the information and tools available? There are, in academic fields as elsewhere, much of conventional approval and hence much of continued inertial work, along lines that were initially thought fruitful.

In economics specifically, is there significant market failure? The relative decline of economic history and other empirical fields, and the high prominance of non-empiricism, are notable. One editor wrote me, a bit defensively, that he was limited in what he could publish by the kind of papers submitted. But of course the topics worked on, and the papers sent in, reflect what is publishable, what will attract research support and what will elicit appointment, promotion, and honor from collegues and seniors, in addition to personal predilections.

Fine points of logic are apt to win approval within the guild. Is enough attention being given to the most pressing economic problems, and to their—often interdisciplinary and concretely factual—solutions? Are economists tempted to sit overmuch in offices and think, supported by basic principles, rather than trudge the often privately unrewarding empirical highways and byways?

The conspicuous relative and absolute state of academic economics suggests to me that it is falling short of the social performance it can reasonably achieve.

■ *I am grateful for support from the East-West Center at the University of Hawaii, and for helpful suggestions from Arthur Goldberger, Lee Hansen, and Burton Weisbrod.*

References

Leontief, Wassily, "Academic Economics," *Science*, July 9, 1982, *217*, 104–107.

"The Naked Economist," *The Economist*, July 17, 1982, 67.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 165-179

Nobel Laureate

On James Buchanan's Contributions to Public Economics

Thomas Romer

When the 1986 Nobel Prize in economic science was awarded to James Buchanan, there was the customary flurry of media attention. As usual, the news stories attempted to summarize the laureate's lifetime accomplishments in a single phrase. In Buchanan's case, it appeared that he had won the prize for figuring out that politicians are motivated by self-interest. As with the "people want to save for their retirement" characterization of Franco Modigliani's research the previous year, the bemusement of the commentators conveyed the impression that there must have been more to it than that.

What the Nobel committee recognized in making its award to Buchanan was his central role in the gradual transformation of the way economists and political scientists study governments and their relationship to the governed. This transformation is still in its early stages. The analytical approach it implies—often labelled "public choice," though I prefer the term "political economy"—is probably not yet the dominant mode even in public economics (and certainly not in political science).[1] But because it provides a logically compelling way to integrate economic and political processes, it has attracted growing attention.

In this essay, I will focus on Buchanan's contributions to public economics, especially his linking of economic and political concerns. Although these topics are

---

[1] Even in relatively new fields, doctrinal distinctions emerge. "Public choice," as I intend to use the phrase, refers primarily to the "Virginia school" approach, especially that of Buchanan and of Gordon Tullock. "Political economy" (in this context) draws on this approach, but leans somewhat more heavily on insights from social choice and game theory. It is also more mathematically oriented in its modelling style. For an entertaining and informative account of some of the varieties of public choice, see Mitchell (1988).

■ *Thomas Romer is Professor of Economics and Political Economy, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania, and Visiting Scholar at The Hoover Institution, Stanford University, Stanford, California.*

Table 1

Works by James M. Buchanan cited in this essay, listed in chronological order.

[1] "The Pure Theory of Government Finance," *Journal of Political Economy*, 1949, *57*, 496–505.
[2] "Federalism and Fiscal Equity," *American Economic Review*, September 1950, *40*, 583–99.
[3] "Federal Grants and Resource Allocation," *Journal of Political Economy*, June 1952, *60*, 208–17.
[4] "Social Choice, Democracy, and Free Markets," *Journal of Political Economy*, 1954, *62*, 114–23.
[5] "Positive Economics, Welfare Economics, and Political Economy," *Journal of Law and Economics*, 1959, *2*, 124–38.
[6] *Fiscal Theory and Political Economy*, Chapel Hill: University of North Carolina Press, 1960.
[7] "Simple Majority Voting, Game Theory, and Resource Use," *Canadian Journal of Economics and Political Science*, August 1961, 337–48.
[8] *The Calculus of Consent* (with Gordon Tullock), Ann Arbor: University of Michigan Press, 1962.
[9] "The Economics of Earmarked Taxes," *Journal of Political Economy*, October 1963, *71*, 457–69.
[10] "What Should Economists Do?" *Southern Economic Journal*, January 1964, *30*, 213–22.
[11] "An Economic Theory of Clubs," *Economica*, February 1965, 1–14.
[12] *Public Finance in Democratic Process*, Chapel Hill: University of North Carolina Press, 1967.
[13] "Efficiency Limits of Fiscal Mobility: An Assessment of the Tiebout Model" (with Charles J. Goetz), *Journal of Public Economics*, 1972, *1*, 25–43.
[14] "Political Constraints on Contractual Redistribution" (with Winston C. Bush), *American Economic Review*, May 1974, *64*, 153–57.
[15] "Public Finance and Public Choice", *National Tax Journal*, December 1975, *28*, 383–94.
[16] "Taxation in Fiscal Exchange," *Journal of Public Economics*, July-August 1976, *6*, 17–29.
[17] "Towards a Tax Constitution for Leviathan" (with Geoffrey Brennan), *Journal of Public Economics*, December 1977, *8*, 255–73.
[18] "Tax Instruments as Constraints on the Disposition of Public Revenues" (with Geoffrey Brennan), *Journal of Public Economics*, June 1978, *9*, 301–18.
[19] *Liberty, Market, and State*, New York: New York University Press, 1985.
[20] "The Constitution of Economic Policy," *American Economic Review*, June 1987, *77*, 243–51.
[21] "Tax Reform as Political Choice," *Journal of Economic Perspectives*, Summer 1987, *1*, 29–35.

generally considered to be the area of Buchanan's major contributions, he has also done considerable work on the theory of opportunity cost and has written extensively on philosophical issues, especially ethics.[2] Within the topics of public economics, I have made a personal, but I think not too idiosyncratic selection. Table 1 lists the works by James Buchanan cited in this essay, presented in chronological order. Throughout the paper I will refer to these articles by number, while all other references will follow the customary author (date) system.

At the time Buchanan began his professional career, positive analysis of the public sector consisted almost entirely of the study of tax incidence. Indeed, Alfred Marshall viewed incidence analysis as a perfect vehicle for teaching much of microeconomics, through the investigation of the impacts of exogenous fiscal variables.[3] (Incidence analysis, with technical advances that permit sophisticated general equilibrium computations, continues to form the bulk of published work in public finance.)

---

[2] Borcherding (forthcoming) and Rowley (forthcoming) discuss these works.

[3] Buchanan [15] makes this point in his discussion of the connections between public choice and public finance.

Normative public finance was based on the (often tacit) assumption that the society is ruled by a benevolent despot. In choosing policies, this ruler acts in "the public interest," correcting market failures, adopting tax structures that minimize deadweight losses, and engaging in appropriate redistribution. Much normative analysis in public economics is still characterized by this view of government.

Buchanan set out to challenge both the positive and normative aspects of the traditional analysis. He did not reject the relevance of positive incidence analysis, but he argued that it was incumbent on economists to study government as an integral part of the economy [6, p. 3]: "Governmental decisions are no more exogenous than are those made by the private economy. ...Any approach to a complete or satisfactory treatment of the public sector must examine as a central feature the way in which collective decisions are made." Since the government is part of the economy, in Buchanan's view, it must not be sectioned off as some exogenous cause of economic events.

Bringing collective decision-making to the forefront of the study of the public sector also requires a shift of normative focus. "The government" is no longer viewed as a machine that computes optimal solutions to social welfare maximands. Instead, in Buchanan's analysis, governments are part of the complex web of institutional arrangements whereby individuals seek to satisfy individual wants. The central normative question is how well the institutional structures perform in doing this.

In this view, the study of the public sector must be conducted within the same analytical framework that economists bring to the study of the private economy. Buchanan insisted on carrying out the positive analysis of government by applying methodological individualism (basing the analysis on individual constraints and incentives) to collective decisions, just as modern economics was doing in the study of markets. Institutions such as "markets" and "governments" do not have objective functions. Rather, institutions emerge through the attempts by individuals or groups of individuals to attain certain goals. In turn, of course, institutional arrangements create opportunities for and impose constraints on individuals.

Early Work

Buchanan's work has displayed a remarkable consistency of theme and outlook over his career. The emphasis on the individualistic view of the state, the focus on government as an exchange nexus, the concern over constitutional rules and their practical outcomes, the influence of the work of Knut Wicksell—these are all present from his earliest writings.

In one of his first publications, Buchanan [1] argued that the government is part of an exchange nexus. Individuals pay taxes and consume government goods and services. More broadly, individuals also rely on the public sector to define the constitutional framework under which commitments and obligations operate. Applying his argument to the study of fiscal incidence, Buchanan pointed out that the then almost exclusive concern with tax incidence tended to obscure the role of government in the exchange process. He stressed the need to look at the total impact of government on an individual by examining the incidence of both taxes and expenditures, rather than that of taxes alone.[4]

A 1950 paper [2] applied the notion of fiscal exchange to the question of horizontal equity in a federal system. Buchanan argued that the norm of "equal treatment of equals" means that, for a given individual, the net effect of tax and expenditure decisions of all levels of government taken together should be independent of the person's geographical location. To accomplish this, he proposed a system of intergovernmental fiscal adjustment. This plan would require geographically discriminatory federal taxes, because equalization would be across individuals rather than among governmental units. Buchanan recognizes that in the United States such discrimination would face major constitutional hurdles, whereas geographically discriminatory federal expenditures are not generally unconstitutional. He considers the legal asymmetry to be due in large part to "doctrinal errors made in economic and fiscal theory which have caused the expenditure side to be treated as a less important area of study than the tax side." We have here an early indication of concern that inappropriate ways of looking at the role of government (in this case, not linking taxing and spending) lead to unsatisfactory rules (asymmetric legal treatment of taxation and expenditure).

This paper, together with a follow-up [3], also anticipates the fiscal equilization debates of the 1960s and later,[5] and makes several important points about the distinction between grant-in-aid programs for specific public services (for example, to assure minimum education standards) and intergovernmental transfers that equalize individual opportunities across governmental units. Interarea transfers are justified not as "subsidization of poor areas," but through the application of horizontal equity among persons, taking into account both tax and expenditure effects.

As Buchanan has written many times, the most powerful influence on his thinking about the analysis of the public sector was the work of Knut Wicksell (see, for example, the Nobel lecture, published in [20]).[6] Wicksell's insistence that attention to the political process be kept part of the study of public finance was to inform all of Buchanan's research in public economics. Buchanan spent the decade of the 1950s in assimilating and elaborating the Wicksellian approach, both in print and through his influence in building a group of scholars at the University of Virginia.

---

[4] In a later look at incidence analysis [6, ch. 7], Buchanan compares direct and indirect taxes, and shows that relative incidence depends in an important way on what assumptions one makes about the use to which the government puts the tax revenues.

[5] For the most part, these debates have focused on the disparities in the abilities of local governments within a state to finance public services (especially education) from local tax bases. The design of a system of grants from higher to lower governments to help offset these disparities has spawned a large literature.

[6] Buchanan's translation of part of Wicksell's 1896 *Finanztheoretische Untersuchungen* is published as "A New Principle of Just Taxation" in Musgrave and Peacock (1958). Although he is not the only modern public finance scholar to note Wicksell's contribution (Musgrave had written an early critique of the voluntary exchange model and recognized Wicksell as a leading light), Buchanan is certainly the person who most thoroughly incorporated Wicksell's approach to the study of public economics.

James M. Buchanan

The publication of Arrow's seminal book on social choice theory (1951) prompted Buchanan to examine the basis of majoritarian decision rules from the Wicksellian perspective. In commenting on Arrow's Possibility Theorem, he argued [4] that the pervasive instability of majority rule—the strong likelihood of voting cycles—is precisely what makes majority rule viable in a heterogeneous democracy. The possibility that, in general, majority coalitions are unstable guarantees that majorities on any particular issue will be temporary. By reducing the probability that any one coalition will dominate for long, this instability can strengthen the willingness of diverse individuals to accept majority rule as a way of making collective decisions. Buchanan was writing two decades before the literature on agenda control, so it is not surprising that his argument does not address the broad scope for manipulation that the instability of majority rule makes possible. Nonetheless, his central point is original and important. Normative evaluation of collective decision rules must be focused not narrowly on the results of the rules, but on how the rules perform in the larger framework in which rules themselves are endogenous. The paper's main message is the need to keep clearly in mind the distinction between the operation of majority voting or any other collective decision rule and the decision about the choice of the rules. This message underlies nearly all of Buchanan's writings from this point on.

The Calculus of Consent

A much more complete elaboration of this theme is provided in *The Calculus of Consent* [8], written with Gordon Tullock. This book is widely considered to be among the most influential publications in modern political economy. Its scope is broad, its goals ambitious. Addressed as much to the political science audience as to economists, it begins with an aggressive development of the validity of methodological individualism applied to collective choice. The tone is combative, and the reader knows early on that the authors are looking for a good fight. They reject the then standard assumption in which the analysis of collective choice was grounded—that agents gather to find some "public interest" or "common good." They write [8, p. 20]:

> No one seems to have explored carefully the implicit assumption that the individual must somehow shift his psychological and moral gears when he moves between the private and social aspects of life. We are, therefore, placed in the somewhat singular position of having to defend the simple assumption that the *same* individual participates in both processes against the almost certain onslaught of the moralists.

While rejecting the notion that the collectivity is the decision unit, Buchanan and Tullock are sensitive to the fact that private and public choice environments are quite different, especially in their informational and incentive aspects. Because of these differences, we may expect individuals to make different exchanges in private than in collective choice.

If collective choice is to be viewed as part of the exchange nexus in which individuals participate, then what is to be kept in the private sector and what should be subject to collective decision? For public allocation, what should be the decision rule? These are the two major questions that Buchanan and Tullock begin to answer. For them, the key to both lies in viewing the economy as a set of institutions whose structure is determined in part through the voluntary participation of individuals. The "contract theory of the state" developed in *The Calculus of Consent* moves Wicksell's analysis of taxation into the realm of constitutional choice. In his Nobel lecture [20], Buchanan noted that he thought this was the book's major contribution. In this formulation, individuals make a "constitutional" decision about how to draw the line between collective action and private action. Individuals recognize that some mutually beneficial exchanges cannot occur if all action is restricted to purely private transactions, and thus set up a constitution to provide political markets (for collectively provided goods) as a way to try to attain such outcomes.

The central problem for individuals at the stage of considering various constitutions is that any collective decision rule other than unanimity will have coercive aspects. How should an essentially voluntaristic model of politics deal with this problem? The answer for Buchanan and Tullock lies in viewing the formulation of a constitution as occurring in an idealized setting, where the following *Gedanken* experiment is performed.[7]

Imagine a group of persons gathered to enact rules to govern future collective decisions. They may foresee that different rules will have different distributional consequences, and so there will be conflicting interests about what rules to adopt. But suppose that they recognize that the rules they are selecting are to apply far into the future—sufficiently so that everyone's expectations about future positions are essentially the same. Or alternately, just assume directly that people at this stage make decisions under a "veil of ignorance" about their own future positions. Then it is more likely that there will be consensus on general rules for collective choice. Indeed, Buchanan and Tullock argue that unanimity over rules may well emerge. In such a setting, they write [8, p. 77], "[A]greement seems more likely on general rules for collective choice than on the later choices to be made within the confines of agreed-on rules." This unanimity argument allows Buchanan and Tullock to sidestep the infinite regress problem of "what rule should be used to decide how to choose the rules," and so on.

---

[7] At about the same time, the philosopher John Rawls (1971) was independently developing a theory of justice using a formulation of the constitutional environment quite similar to that of Buchanan and Tullock.

In deciding what activities to undertake through collective organization, the main concern of individuals considering possible constitutions is to minimize what Buchanan and Tullock call "costs of social interdependence." These consist of direct external costs and the costs of organizing to mitigate the externalities. In some cases, private organization through voluntary exchange will suffice.[8] In others (such as pure public goods provision), collective organization is needed. In assigning some role to collective action, a decision must also be made as to the voting rule required for collective decisions.

If collective decisions require approval by only a relatively small minority, organization or decision costs are kept low. Such a rule, however, implies a high probability that external costs are borne by the majority who are on the losing side. At the other extreme, under unanimity, external costs are fully internalized (since unanimous consent cannot be achieved without compensation), but decision costs can be very high. Weighing these typically opposing costs is the essence of the constitutional calculus facing each participant.

Where the potential externality costs are likely to be high relative to the decision costs, such as in the definition of basic property rights, the constitutional choice is likely to favor a near-unanimity rule. For other issues, less stringent rules will suffice. If we assume that rules emerge in this way, the Buchanan-Tullock approach becomes the basis for a positive theory of institutional structure.

In *The Calculus of Consent*, several chapters are devoted to analyzing the allocative effects of majority rule.[9] (These chapters are based in part on [7].) A simple game-theoretic example demonstrates that in voting situations that, though abstract, do have some realism, the likely outcome is a relative "overprovision" of collectively supplied goods. From this perspective, it follows [8, p. 169]:

---

[8] Ronald Coase (1960) explored private arrangements for mitigating externalities. Coase was concerned primarily with small groups. His analysis dovetails with that of his Virginia colleagues Buchanan and Tullock, who focused on larger groups.

[9] The economic aspects of using voting as a way to allocate resources had been explored by Hotelling (1929) and Bowen (1943). Hotelling drew the analogy between locational equilibrium among competing firms and competition among political candidates. Also, competition between two political parties is the basis for much of the discussion in the pathbreaking work of Downs (1957). Bowen analyzed voting as a way to decide spending on a collectively provided good. Bowen noted that, for some special cases, voting could lead to a Pareto optimal allocation of the publicly provided good.

> There is nothing inherent in the operation of [a majority voting] rule that will produce "desirable" collective decisions, considered in terms of individuals' own evaluations of possible social alternatives. Instead, majority rule will tend to result in an overinvestment in the public sector when the investment projects provide differential benefits or are financed from differential taxation. There is nothing in the operation of majority rule to insure that public investment is more "productive" than alternative employments of resources.

This result is independent of the social choice properties of majority rule discussed by Arrow. Instead, it follows from the fact that the majority can shift some of the cost of provision onto a minority.[10] The private marginal cost of public provision to someone in the majority is less than the social marginal cost. (This cost shifting is one example of the sort of "external cost" that individuals may anticipate in the constitutional stage.) This, in turn, leads to "overprovision" of this particular good. Some of these effects may be mitigated if, through vote trading across issues, voters are effectively able to express intensity of preference as well as simple approval/disapproval.

It is also possible to construct examples where majority rule leads to underprovision of the public good. Although Buchanan and Tullock seem to view this outcome as less likely, one of the central messages of *The Calculus of Consent* is that, in any case, the usual Pareto criteria for evaluating allocations may not be the right way of looking at things. Indeed, for much of Buchanan's later work, it is the normative elaboration of the constitutional argument that has been most important. The normative argument advanced in *The Calculus of Consent* develops points made earlier in [1], [4], [5], and [7]. In a nutshell, the argument is that looking at the efficiency of allocations after they have arisen from the operation of particular rules may often be missing the point. One should look, instead, at the process that gave rise to the rules, and examine the rules themselves in light of that process.

After recognizing that majority rule is likely to lead to inefficient outcomes, for example, one should look back at the origins of the rule as a decision-making device. When they were considering their constitution, individuals may well have anticipated at least some of the costs of using majority rule, including the possible overextension of public relative to private provision. These costs would have been traded off against the costs of complete privatization as well as the costs of using some other, more restrictive, form of voting. From the perspective of those deciding on a constitution, the rule selected may have been the efficient choice—even though it may appear inefficient after being put into effect. Buchanan and Tullock write [8, p. 210], "To make normative statements concerning whether or not governments undertake 'too much' or 'too little' activity seems to be rather wasted effort unless one is prepared to suggest some possible [i.e., feasible] modifications in the organizational rules through which decisions are made..."

---

[10] This possibility may be limited by the ability of members of the minority to leave the polity.

In this normative setting, the performance of the public sector cannot be evaluated as if the economist were giving advice to a benevolent despot. Instead, one must bear in mind the complex process of political exchange from which collective decisions result. This does not make normative analysis vacuous, but it does make it more difficult. As Buchanan later explained [15, p. 391], it requires accepting a "paradigm of the political world... that embodies the democratic-individualist standard that persons should get what they want so long as each person counts for one." So changes are in order if policies result from failures of representation or information that distort individual choices or votes.

In focusing on the distinctions between optimality from the perspective of those designing the constitution and optimality from the perspective of those living under it, the constitutional criterion anticipated the work of theorists who began looking at the design of allocation mechanisms in the 1970s. That literature also stresses the importance of evaluating institutional arrangements by looking at how well they serve to facilitate exchange in the light of informational considerations and organization costs, rather than judging only by the efficiency of particular allocations that result from a particular institutional choice.[11]

Applying *The Calculus* to Public Finance

Earmarked and General Fund Taxes

The apparatus developed in *The Calculus of Consent* was applied by Buchanan to a wide variety of questions in public finance. A particularly interesting example is his analysis of the economics of earmarked taxes in [9] and chapter 6 of [12]. Earmarking ties the use of a particular tax to a specific type of expenditure; for example, federal taxes on the sale of gasoline are used to finance expenditure on highways. Traditional analyses of earmarking tended to decry the practice as making little budgetary sense, serving only to reduce the flexibility of administrators in their attempts to allocate revenues among functions.

Buchanan approached earmarking by considering how the method of financing government activities affects the demand for each activity. If individuals differ in their demands for government services, then they will also differ in their preferences over the method of financing. Thus, some groups will prefer that services be financed from general revenues, while others would like to see them "unbundled," with taxes earmarked for particular services. The political equilibrium depends on the distribution of preferences. In this view the tax structure becomes a matter of importance, because it will affect the level as well as the mix of expenditures.

Buchanan also argued that tax structures have important informational effects. As a close student of the Italian school of public finance[12]—the school whose adherents had developed the notion of fiscal illusion—he noted that efficient individual decision-making requires awareness of costs as well as benefits of public activity. Tax structures that are divorced from particular expenditures are more likely to foster ignorance of the connection between costs and benefits. Or as he put it [18, p. 317], "General-fund financing insures that fiscal choices are made under almost maximal uncertainty."

---

[11] An early statement can be found in Hurwicz (1960), but its fuller development occurred later. See especially Hurwicz (1972). For an exposition in the context of contract theory and mechanism design, see Holmstrom and Myerson (1983).

[12] See ch. 2 in [6].

Clubs and Communities

It is often convenient to divide goods as being either purely private or purely public. But in his 1965 paper, "An Economic Theory of Clubs," Buchanan [11] began the investigation of the organization of exchange for goods that do not fall nicely at either extreme. Such goods have the property that, unlike pure private goods, some sharing may be desirable. Unlike pure public goods, however, consumption of these goods may be subject to congestion, so at some point exclusion may be desirable (crowding effects occur). Again Buchanan based his analysis on individual utility maximization. For the goods in question, utility depends not only on the individual's consumption, but also on the number of people sharing the good. He developed first-order conditions for efficient allocation of "club goods," and showed that optimal organization of "ownership-consumption" rights may be analyzed with the tools of neoclassical demand theory.[13]

One area where the theory of clubs has been applied is in the study of community formation and organization. Some of this work has been done in the context of trying to formalize the model suggested by Tiebout's influential article (1956) on the efficient provision of public goods. Tiebout suggested that since individuals can move between different local government jurisdictions, local governments can be considered as a market offering a greater or lesser range of local public goods at greater or lesser cost. Competition among jurisdictions would lead to efficient provision of local public goods.

In one of the first papers to investigate the efficiency properties of the Tiebout framework, Buchanan and Goetz [13] focused on the adjustment process inherent in "voting with one's feet." When public goods and services are locally provided, and communities are geographically defined, a person moving between communities imposes fiscal externalities on both the place he leaves and the one he enters. This is because, in making locational decisions, individuals compare only private costs and benefits. They do not consider the impact of their locational decisions on the tax prices of other residents.

For these externalities to be internalized, Buchanan and Goetz pointed out, communities would have to adopt discriminatory tax and spending structures and entry rules that capture individual-specific locational rents. A system of "profit-maximizing" towns could, in principle, accomplish this. However [13, p. 38], "the organization and operation of a fiscal sharing group on this basis violates the central notion of free migration, the notion upon which the models of Tiebout adjustment process are initially founded." Since the Buchanan-Goetz paper, a variety of other difficulties with the Tiebout claim have been identified; efficient allocations arise only in quite special cases.[14]

---

[13] Among other things, this analysis showed that the pattern of sharing arrangements will be related to such economic factors as income [11, p. 1]: "European hotels have more communally shared bathrooms than their American counterparts. Middle and low income communities organize swimming-bathing facilities; high income communities are observed to enjoy privately owned swimming pools."

Redistribution

In the constitutional framework developed in *The Calculus of Consent*, the redistributional consequences of voting rules are incorporated as part (perhaps the major part) of the externality component of interdependence costs. The contractually derived institutions may incorporate explicit redistributive norms. These norms can reflect agreed-upon principles of justice that call for income or wealth transfers among individuals in the post-constitutional period.

In an insightful brief paper, Buchanan and Bush [14] pointed out the limitations that post-constitutional politics impose on the application of constitutionally derived redistributive rules. For example, suppose that the constitution embodies the Rawlsian difference principle. By this principle, differences in individual incomes are justified only to the extent that the income distribution that results permits a higher level of well-being for the worst-off individual than could be achieved under alternative distributions. Rawls (1971) argues for this principle by invoking the idea of a "veil of ignorance;" if people do not know what position they will hold in society, Rawls claims, they will take care to assure that the worst-off individual is as well off as possible. Implementation of this principle will typically require policies to redistribute income in post-constitutional periods.

But Buchanan and Bush point out that when the redistribution is actually scheduled to take place, people are no longer operating under the "veil of ignorance." At that time, all people know their own incomes and have some knowledge of how proposed transfers will affect them. Since the redistribution must occur through the political process, it must be acceptable to the constitutionally required majority. It is easy to show that, in many plausible cases, a democratic political process will not implement the constitutionally desired redistributive norm.

While recognizing such limits on the practicality of placing redistributive concerns in a constitutional setting, Buchanan has insisted on the normative message that the constitutional level is precisely where debate about redistribution belongs. Otherwise, the day-to-day operation of government becomes a struggle over redistribution, with more and more frequent instances of majorities exploiting minorities. The constructive vision of politics as an exchange process for mutually advantageous trades is, in this view, eroded by use of the public sector as a battleground for redistribution.

Of course, in practice, it will be very difficult to separate out all questions of distribution from those of allocation, and Buchanan does not provide unambiguous answers on how to do this. Rather, he emphasizes the importance of continually scrutinizing the legitimacy of government action in the context of voluntary exchange. An extended quote is perhaps the best way to summarize the forcefulness and asperity with which Buchanan [16, p. 29] holds this view:

---

[14] See Stiglitz (1977) and Bewley (1981).

> The exchange framework tends to promote a constructive attitude toward governmental process, an attitude that accentuates the cooperative aspects, that underlines the prospects for mutuality of gain for all citizens. The alternative framework [advocating government as the agent of ongoing redistribution, for example, along the lines of the optimal income tax literature] may lead citizens and their political spokesmen to accentuate the profit-and-loss aspects of political competition, to promote a willingness on the part of a dominant coalition to impose its will on its minority opposition, and conversely, to generate in the minority an acceptance of a quasi-Marxist and exploitative view of governmental process. These public attitudes become more significant as the size of the public... sector grows relative to the national product. To this point in American history, the relatively limited scope of overt fiscal exploitation can, I think, largely be explained by a generalized and widely-shared sense of fiscal exchange which has informed public thinking about both constitutional and post-constitutional decision processes. To the extent that the fiscal process, and politics generally, comes increasingly to be viewed as a source for profit opportunities, unrestricted by constitutional precepts, we must predict decreasing fiscal equity along with further departures from efficiency, almost regardless of how these objectives are defined.

The Supply Side of the Public Sector

The framework developed in *The Calculus of Consent* brings politics into the purview of public economics. The view of politics is, however, entirely driven by the demands of individual citizens. There are no political agents or government bureaucracies. To the extent that representative government is discussed, the focus is on the appropriate degree of representation, taking into account gains in reducing decision costs as well as possible losses due to incomplete representation of individual preferences. The incentives of representatives and of suppliers of public services are absent from the picture.

In this setting, the actual operation of the government may not be a social-welfare-maximizing machine, but it is still very much a black box. The level of public services (taxes, expenditures) is determined entirely by demand considerations, as aggregated through the constitutionally specified voting rules.[15] The supply side of the public sector delivers these services at the level demanded. However, a growing body of work in economics and political science had been looking at the behavior of legislators and bureaucracies. Particularly influential was the work of the "Rochester school" of political science on legislative politics (Mitchell, 1988), and books by Gordon Tullock (1965) and William Niskanen (1971) on bureaucracy.

---

[15] In public economics, this is most frequently seen in the by now standard application of "median voter" characterizations of political/economic interaction.

Buchanan's published work pretty much ignored the supply side until about the mid-1970s. Prompted, I think, largely by the rapid growth of government spending in the late 1960s and early 1970s, he began to look at the problem of how supply side considerations should be brought into the constitutional choice framework.[16]

Key ingredients in this endeavor were his acceptance, at least at the analytical level, of Niskanen's view of government bureaucracies as seekers of ever larger budgets, subject to the constraints of the political process. Niskanen provided informational and structural reasons why these constraints were not likely to eliminate or perhaps even significantly curb the tendency toward the budget-maximization. In work with Geoffrey Brennan ([17] and [18]), Buchanan asked: Suppose that, when the constitution is being drawn up, individuals forecast that in post-constitutional periods, governments will act in a revenue-maximizing, rather than a purely demand-driven way. What constitutional rules might serve to mitigate this? (The analysis requires, of course, that the government will obey constitutional rules.)

The answer that emerges is to choose rules that limit the revenue-raising ability of government and constrain its freedom to choose the spending mix. One consequence of this view is that expanding the comprehensiveness of the tax base is not generally a good idea; it may facilitate wasteful increases in taxes. We also see a rationale for designing a structure of earmarking so that each type of government activity is financed by a tax on a base that is complementary to the activity. Earmarked taxes may result in efficiency costs when viewed from the time when the constitution is already in place, but from the standpoint of constitutional design, they serve the important purpose of constraining excessive spending. (In his article in the Summer 1987 issue of this journal, Buchanan [21] applies this approach in a critique of the 1986 Tax Reform Act.)

In his most recent work, Buchanan has focused largely on elaborations of these considerations, particularly with respect to the use of rules that mitigate prisoners' dilemma problems that arise from the rational, self-interested behavior of political entrepreneurs. Most prominent among these is his characterization of the tendency toward large budget deficits as a result of the incentives faced by legislators, and his advocacy of constitutional rules requiring balanced budgets.

Concluding Thoughts

In all of this, Buchanan has remained steadfastly controversial. Though his analytical work is based firmly on the calculus of individual choice, his discussion of political organization has always had a moral concern. He has inveighed against Keynesians for undermining what he believes was a social contract against routine deficit spending. He is equally harsh on the Reagan administration for further eroding the moral consensus against burdening future generations with the excesses of today's deficits. He has taken mainstream economics to task for being driven by technique rather than substance, and has decried what he sees as the sterility of most theoretical work. He is no kinder to the bulk of empirical research, and as far as I know has never published an empirical paper. I suspect he even views much of the current work in political economy/public choice, as it becomes more fully integrated into the mainstream of economics or political science, with some dismay.

---

[16] Buchanan gives his own views on his change of focus over this period in ch 2 of [19].

His message to economists in his Nobel lecture [20] is not much different from that in his 1964 Southern Economic Association presidential address [10]: In proffering policy advice, economists should focus on the underlying institutional structure, and keep the interaction of economics and politics firmly in mind. This is particularly good advice for students of public economics. The integration of politics and economics into a modern, coherent framework has barely begun. Throughout his work, Buchanan has championed the basic importance of this research and played a central role in defining its agenda.

■ *I would like to acknowledge helpful conversations with Thomas Borcherding, Otto Davis, and Howard Rosenthal. Allan Meltzer, Peter Van Doren, and the editors provided valuable comments on an earlier draft. Borcherding deserves my special thanks for making available to me draft versions of his own survey of Buchanan's work, including Borcherding's (1987) personal and witty remarks prepared for the AEA luncheon honoring Buchanan. (Due to the vagaries of Chicago winter weather, the remarks were actually delivered, with ad lib embellishments, by Charles Plott.) Borcherding (forthcoming) contains an extensive, though not exhaustive, bibliography of Buchanan's oeuvre.*

References

Arrow, Kenneth, *Social Choice and Individual Values*, New York: John Wiley and Sons, 1951.

Bewley, Truman F., "A Critique of Tiebout's Theory of Local Public Expenditures," *Econometrica*, May 1981, *49*, 713–40.

Borcherding, Thomas E., "In Honor of James M. Buchanan, Nobel Prize Winner in Economics for 1986" Presented by Charles Plott at American Economic Association Meetings, Chicago, December 29, 1987.

Borcherding, Thomas E., "The Economics of James M. Buchanan," paper presented at Southern Economic Association Meetings, Dallas, Texas, November 25, 1985. Also, forthcoming in *Economic Inquiry*.

Bowen, Howard R., "The Interpretation of Voting in the Allocation of Economic Resources," *Quarterly Journal of Economics*, 1943, *58*, 27–48.

Coase, Ronald, "The Problem of Social Cost," *Journal of Law and Economics*, 1960, *4*, 386–405.

Downs, Anthony, *An Economic Theory of Democracy*, New York: Harper, 1957.

Holmstrom, Bengt, and Roger Myerson, "Efficient and Durable Rules with Incomplete Information," *Econometrica*, 1983, 1799–1820.

Hotelling, Harold, "Stability in Competition," *Economic Journal*, 1929, *39*, 41–57.

Hurwicz, Leo, "Optimality and Informational Efficiency in Resource Allocation Processes." In Arrow, Kenneth, S. Karlin, and P. Suppes (eds.), *Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press, 1960.

Hurwicz, Leo, "On Informationally Decentralized Systems." In Radner R., and C. B. McGuire, eds., *Decision and Organization*. Amsterdam: North Holland, 1972.

Mitchell, William C., "Virginia, Rochester, and Bloomington: Twenty-Five Years of Public Choice and Political Science," *Public Choice*, February 1988, *56*, 101–19.

Musgrave, Richard A. and A. T. Peacock, eds. *Classics in the Theory of Public Finance*, London: Macmillan, 1958.

Niskanen, William, *Bureaucracy and Representative Government*, Chicago: Aldine, 1971.

Rawls, John, *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.

Rowley, Charles, "Philosophical Writings of James Buchanan," *Economic Inquiry*, forthcoming.

Stiglitz, Joseph E., "The Theory of Local Public Goods." In Feldstein, M. S. and R. P. Inman, eds., *The Economics of Public Services*. London: Macmillan, 1977.

Tiebout, Charles, M. "A Pure Theory of Local Government Expenditures," *Journal of Political Economy*, October 1956, *64*, 416–24.

Tullock, Gordon, *The Politics of Bureaucracy*, Washington: Public Affairs Press, 1965.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 181–189

Puzzles
Blockades, Carrier Missions, Secret Intelligence, and More

Barry Nalebuff

As usual, this feature begins with several speed puzzles; answers can be found at the end of the problems. Following is one longer puzzle for which readers are invited —nay, challenged—to submit their own answer. The puzzles in this issue focus on everything from diversifying strategies in naval warfare to ordering in restaurants to finding the quickest way around Delhi. The column ends with reader mail, including submitted solutions to Schelling's Poison Pill and the fairest way to score a contest.

Please send your answers, comments and favorite puzzles to me directly: Barry Nalebuff, "Puzzles," Department of Economics, Princeton University, Princeton, NJ 08544. Good luck.

Puzzle 1: Sunk Costs

Jack Hirshleifer (UCLA) writes in with the following puzzle: "This one is an almost excessively simple problem in terms of algebra. The puzzle is to explain the seemingly counterintuitive result.

A naval procurement decision involves a choice between two options: large carriers which will operate singly, or small carriers which will operate in pairs. A pair of small carriers costs as much as one large one in all relevant dimensions. The difference between the two options is on the score of vulnerability. Under enemy attack, a large carrier has a probability p of surviving and executing its mission, while each of the smaller carriers has probability $p/2$. However, the small carrier pair can carry out its mission *even if only one survives*. (Assume that for each member of the pair,

■ *Barry Nalebuff is Assistant Professor of Economics, Princeton University, Princeton, New Jersey.*

the risks are independent.) If we are only interested in execution of the mission, which option is superior?"

Puzzle 2: Cheaper by the Half-Dozen

Richard Zeckhauser (Harvard-JFK School) tells of a puzzle from dining in *The Ocean Club* restaurant (located in the Charles Hotel, Cambridge, MA). This restaurant offers half-portions of the main course entrees for less than half the price of the full portion. When Richard and his companions used this option, they discovered that the half-portions were at least half the size of the full portions and more than that, they came with full portions of the accompanying vegetables. As economists, they debated possible explanations. Could it be that the restaurant wanted to encourage people to try more than one dish? If diners most clearly remembered their favorite dish, then this incentive to sample would increase the chance that someone would find a "winner" and this would lead to great word of mouth advertising. Or did this half portion ordering reduce the variance in the distribution of orders so that there would be less spoilage and hence lower costs? To settle the dispute, the maître d' was called over. What was his explanation? Hint: No one at the table guessed correctly.

Puzzle 3: The Chicken or the Egg?

Emmett Keeler (Rand) offers the following little puzzle on estimation to get at stock versus flow issues. Question: Are there more hens (the kind of chicken that lays eggs) or broilers (the kind of chicken we eat) alive today in the United States?

To answer this question, you should assume that laying hens and broilers are two different species. You will want to make some back-of-the-envelope estimations for how many eggs and how many chickens are consumed and for the time it takes to produce an egg versus a broiler. Although these numbers may well be outside your area of expertise, the difference in population sizes is a factor of fifteen, so there is plenty of room for rough guesses.

Puzzle 4: War and Profits

Avinash Dixit (Princeton) describes a puzzle from James McPherson's recent best-seller, *The Battle Cry of Freedom*. With the blockade of the South during the Civil War, "Wilmington and Nassau became wartime boom towns.... The chance of profits from a successful voyage outweighed the one in three chance (by 1864) of capture. Owners could make back their investment in one or two round trips, clearing pure profit with every subsequent voyage." Given the returns and the chance of being caught reported by McPherson, was there a potential for profitable entry into this smugglers' market? Remember, smugglers made money and ran the risk of capture on *both* legs of the round-trip from Nassau to Wilmington.

Puzzle 5: The Delhi Dilemma

To rival the best location in Manhattan puzzle, Kaushik Basu (Delhi School) offers his version of a Delhi dilemma. In New Delhi's Connaught Place area, there are two roads which are concentric circles with plenty of radial roads as shown below. All roads are two-way, as shown in Figure 1. In negotiating Connaught Place, drivers in Delhi regularly encounter a small dilemma. While going from one point to another on the outer circle, they face two options: they can drive all the way along the outer circle, or instead, they may move to the inner circle, drive along it and come out at the relevant point on the outer circle.

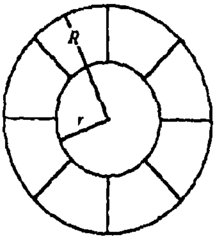


Fig. 1. Connaught Place.

Which route is shorter? To answer this question, you should assume that *there are radial roads everywhere*. If you wish, you may take the radius of the inner road as r and of the outer road as R. How far must one be traveling along the outer road before it becomes worthwhile to take the inner road?

Answers to Puzzles 1–5 follow Puzzle 6.

Puzzle 6: Significance is in the Eye of the Beholder

Once again Jack Hirshleifer comes up with a fascinating puzzle. In World War II, the Allies employed a technique called 'serial number analysis' to estimate military production. For example, if a captured tank of a particular model bore serial number 116, then (assuming away deliberate enemy misperception) at least 116 units must have been produced.

Imagine that, before any captures of a new enemy weapon, we have made plans on the basis of a very reliable espionage report that 1000 units were produced, so that $N = 1000$. Now we learn that the spy's report may have been garbled in transmission, so that there is a 50 percent chance that he may have really reported $N = 10{,}000$. Only two actions are available to us, one being superior if $N = 1000$ and the other if

$N = 10{,}000$. After careful consideration of the costs and benefits, we determine that, in the absence of any additional information, it is best to act as originally planned.

Before this decision is carried out, however, we manage to capture a single enemy unit bearing serial number 992. Using the "tests of significance" approach or the "Bayesian" approach, should we now change our decision?

A statistician working in terms of tests of significance might say: "Clearly the null hypothesis is $H_0: N = 1000$ while the alternative hypothesis is $H_A: N = 10{,}000$. An 'upper rejection region' should be established corresponding to the acceptable risk α of a type I error, that is, falsely rejecting the null hypothesis. In practice, we statisticians use $\alpha = 0.05$ as a rejection criteria, except that when we are being very careful we sometimes use $\alpha = 0.01$ instead. With $\alpha = 0.05$, serial numbers from 951 upward would be in the rejection region; with $\alpha = 0.01$, the rejection region is from 991 upward. Either way, 992 is in the upper rejection region for the null hypothesis, so we should accept the alternative hypothesis."

A Bayesian statistician replies: "That's ridiculous. If the observed serial number is anywhere in the range from 1 to 1000, that increases our confidence in the null hypothesis that $N = 1000$. Since our prior beliefs already favored acting in accord with the null hypothesis, we should obviously continue to accept the null hypothesis."

Which decision is correct? What are the implications for the two approaches to interpretation of statistical evidence? Dale Poirier's article in Winter 1988 issue of this journal offers some guidance—perhaps even the answer to the clever reader.

Hirshleifer offers a hint: "In using the tests of significance approach, it is incorrect to entirely ignore Type II error. But that raises another question: why do practical statisticians, practically all the time in fact accept or reject hypothesis solely in terms of the $\alpha = 0.05$ or $\alpha = 0.01$ convention?"

Answers to Speed Puzzles

Answer to Puzzle 1

Hirshleifer provides his own answer: "Intuitively, since only one member of a pair of small carriers has to survive for the mission to be carried out, the small-carrier option seems to be superior. But algebra indicates this is not the case. The large carrier executes the mission with probability p. For the small carrier, the corresponding probability of success is the chance that not both are sunk: $1 - (1 - p/2)^2 = p - p^2/4 < p$. To explain the seemingly counterintuitive result, notice that the mathematical expectation of carrier survival is exactly the same for each option. But, from the point of view of executing the mission, for the small-carrier pair some of this expectation is 'wasted' in that, with probability $p^2/4$, *both* survive whereas only one is needed."

A second advantage to the large carrier option is the future valuation. Although the expected number of surviving ships is the same (p) in both cases, with the large ship, a single surviving vessel is much more valuable than a small ship. For example, if there are two missions, the chance that the large ship will successfully execute both is p^2 while for the small carriers the chance is $2(p/2)(1-p/2)(p/2)+(p/2)^2(p-p^2/4)=p^2/2-p^4/16$, or less than half the chance with the large carrier option.

Answer to Puzzle 2

The maître d' listened patiently to all the alternatives and rejected each one. The true reason, he explained was that this was an *irrational* thing to do. People noticed this discrepancy quite regularly and told their friends about this inconsistency on the menu. As a result, this generated tremendous word of mouth publicity for the restaurant at very little cost. This puzzle is simply one more feather in their cap. In the spirit of Sherlock Holmes, when all explanations appear to fail, the explanation must be that there is no explanation.

Answer to Puzzle 3

There are roughly 15 times as many broilers as laying hens. A modern hen lays 5 eggs a week, up from 3 eggs a week in the 1920s (see Conniff, 1988). The statistical abstract for 1981 shows egg output at 190 million eggs per day; this implies the average person eats almost an egg a day (including the eggs in cakes, mayonnaise, and so on). The stock of hens needed to satisfy this consumption flow is $(7/5) \times 190$ million or about 265 million hens.

The modern broiler takes only 7 weeks to reach its four pound market weight, down from 16 weeks in the 1920s (Conniff, 1988). Hence the next seven weeks of consumption must all be alive at the present time. Working backwards, we find that the average per capita consumption of chicken is around 2.3 per week. A weekly consumption flow of 2.3 chickens requires a stock of slightly over 16 chickens per person, which is in line with the statistical abstract for 1981 which lists 4.1 billion broilers. While that seems to be a lot of chicken, U.S. consumption of chicken now outweighs beef. Colonel Sanders and Chicken McNuggets are doing their part.

Answer to Puzzle 4

Let K equal the investment in a blockade running ship, including the cost of stocking it with goods. A successful trip leads to net profits of X, which includes the restocking cost at the end of each trip. The probability of capture on each trip is denoted by p. With risk neutral smugglers and free entry, total investment costs should equal expected profits.

With a p chance of capture, the expected number of trips until capture is $1/p$, but on the last trip, no profits are made as the ship and cargo are captured. Hence the number of profitable trips is $1/p - 1$ and this must be enough to offset the literally sunk investment costs: $K = X(1-p)/p$. If $p = 1/3$, then $K = 2X$, and the investment is recouped in one successful *round* trip. For recoupment in two successful round trips, we have $K = 4X$, which needs $p = 1/5$, which is closer to the average figure for the war as a whole (1/6, given later on page 380).

Answer to Puzzle 5

If one could travel all the way to the center of the circle, it would be possible to get to any point along the circumference in distance $2R$: the distance to the center is R and then back to any point on the circumference is another R. This is shorter provided one is travelling more than $2R$ along the circumference, which corresponds to $1/\pi$ of the way around the circle or an angle of more than 2 radians. (A radian is $360/2\pi$ degrees.) More generally, if one is travelling over an angle of θ radians, the distance along the circumference is $R\theta$. If instead, one first travels a distance $R - r$ towards the center this will reduce the circumference portion to $r\theta$ at a cost of $2(R - r)$ (the roundtrip). This is worthwhile provided $2(R - r) < \theta(R - r)$ or $\theta < 2$, the same condition as before.

It is worthwhile noting that since $R - r$ factors out, the result does not depend on the relative distances between the two circular roads. All that matters is what fraction of the way around one is going. As soon as one is travelling more than 2 radians or $360/\pi$ degrees it is worthwhile heading into the center (and doing so as far as possible).

Mail

Comments on Puzzles from Fall 1987

Puzzle 5. Unhealthy Constitution. Tom Schelling told the tale of a fantastically wealthy company which had an unusual company charter. The firm was owned and managed by thirteen shareholders who, each owning 1/13th of the stock, constituted the board of directors. The board had full authority not only to run the company but also to redistribute ownership shares and to determine board membership. All decisions of the board were by simple majority vote. Voting was open, each member voting yes or no in his turn. The turns progressed clockwise around the table. No motion required a seconder. The voting always began with the director sitting to the left of the proposer. Anyone making a proposal was recorded as voting for his proposal.

Among the provisions in the company's charter, one was unusual. It was designed to discourage "constitutional change," which was interpreted as any change in the membership of the board, or any change in ownership of the shares of the company. The rule was that if anyone offered a motion changing the voting rules, changing ownership of shares, or changing board membership, and this motion failed, he would be deprived of his ownership rights and of his membership on the board. The confiscated shares would be divided evenly among the remaining members of the board. Furthermore, anyone who voted in favor of a motion that failed would suffer the same fate as the proposer; namely, he would lose his shares in the company and his membership on the board.

In spite of all these "poison pills" someone proposed a motion which provided, in a somewhat devious way, for confiscating all of the shares of his fellow board members, making him owner of 100 percent of the company and sole member of the board of directors, all in return for a $1 compensation. His motion passed unanimously. What was the motion that received their twelve yes votes?

Tom Romer (Carnegie-Mellon) writes in with a correct solution. "Solving it cost me the better part of a day—but, at least it was the *better* part of the day. It's obviously a backward-induction puzzle that requires a motion with contingencies tied to the vote outcome. There are many variants of the details that would work, but all motions have to have an element like item (1) to make the pivotal voter prefer passage and, obviously, an element like item (2) to deliver unaimity. Mr. Devious (D) makes the following motion:

> 1. If this motion passes but not unanimously, the shares of those voting No will be taken away and divided equally among those, *other than* D, who voted Yes. Those who voted No will receive nothing and be removed from the board.
> 2. If this motion passes unanimously, everyone's shares will be given to D. Everyone other than D will receive $\varepsilon > 0$, and will be removed from the board.

"To verify that this proposal works, we start with the last member to vote. Obviously if there are already 7 or more Yes votes, she will vote Yes (as otherwise she will receive nothing). If the vote is tied at 6–6, then voting Yes gives her a payoff of two shares—her own plus one-sixth of the six shares held by the six No voters. If she votes No, her payoff is only 13/7 shares—her own plus one-seventh of the six shares held by the six Yes voters. So she will vote Yes.

"Now consider the 12th (next-to-last) voter. If there are already six or more Yes votes, then she can predict that the motion will pass, since voter 13 will vote Yes by the above argument. So #12 will vote Yes too. Suppose there are only 5 Yes votes when #12 casts her vote. If she votes Yes, so will #13, and the final outcome will be 7-6 in favor; #12's payoff will be two shares as above. If she votes no, the motion cannot pass, so #13 will also vote No, leading to a 5-8 outcome; #12's payoff is 13/8 shares—her own share plus 1/8 of the five confiscated shares. So #12 will vote Yes.

"We can carry this same argument back down to the eighth voter. If before her vote, 2 or more Yes votes have been cast, she can predict that the motion will pass regardless of her vote, so she will vote Yes. If only the proposer has so far voted Yes, then a Yes vote by 8 will lead to a 7-6 passage and a payoff of 2. Voting No will doom the motion to a 1-12 failure and #8's payoff will be a mere 13/12 shares. So #8 will vote Yes in response to Mr. Devious's proposal.

"For voters #2 through #7, there is really no choice. No matter what they do, #8 through #13 will vote Yes so that passage is assured. Voting No will result in a payoff of 0, while voting Yes guarantees at least $\varepsilon > 0$. So they will all vote Yes. The final result: Unanimous passage or murder (probably serial murders, at that)."

When I gave this problem to my students, Eliot Robin came up with an ingenious alternative solution. One of the features that makes his solution different is the last voters will not necessarily vote to pass the resolution when pivotal. Yet, in equilibrium, the resolution always passes. Mr. Devious makes the following proposal:

1. If someone votes No and the next person votes Yes, then the yes voter gets to keep their own shares plus the shares of the previous No voter.
2. If this motion passes unanimously, everyone's shares will be given to D. Everyone other than D will receive $\varepsilon > 0$, and will be removed from the board.

This proposal works via *pair-wise* backwards induction. Imagine that when it comes to the last two voters there are 6 Yes votes. If #12 says No then #13 will want to say Yes in order to get two shares rather than 13/7ths. If #12 says Yes, then the proposal passes.

Next go back to the #10 and #11 pair. If #10 faces 6 or more Yes votes, he can predict the proposal will pass, so he votes Yes. If there are only 5 Yes votes and he votes No, then #11 will want to vote Yes in order to take #10's shares—he can do so with confidence as this will leave the #12 and #13 pair with 6 yes votes so that one of them will vote Yes and the proposal will pass.

Working backwards, we see that in each of the six pairs [(#2, #3), . . . , (#12, #13)] at least one of the two will always vote Yes provided each of the previous pairs has come up with at least one Yes vote. Since everyone can then predict that the proposal will pass, they will all vote Yes in order to get ε rather than 0.

Although these proposals may seem a little far-fetched, in fact they are a simple extension of the same coercive bid involved in two-tiered tender offers. As described in Grossman and Hart (1980), imagine that the fair market value of a share is \$1.00. Let a raider make an unconditional offer of \$1.01 for the first 50 percent of the shares tendered and zero thereafter. If more than 50 percent of the shares are tendered, the \$1.01 will be prorated; each sharehlder will receive \$1.01 × (50 percent)/(percent tendered). In this case, tendering is a dominant strategy: if the offer fails, tendering is profitable as you gain a penny per share; if the offer succeeds, again tendering is a good idea as you get something as opposed to nothing.

Puzzle 6. Close Counts in Hand Grenades, Ballroom Dancing, and Game Design. This puzzle asked how one can design scoring rules to ensure that two players of equal ability will have an equal chance of winning even when they face differential conditions (for example, in ping-pong the two sides of the table are often uneven). To restore equity, the two players should switch sides sometime during the match. When?

Emmett Keeler (Rand) writes in with the solution: "The most interesting case is when you allow the sum of the scores to be used in determining the switch. Here tennis and beach volleyball do it right, by forcing each team to play half the time on each side. For ping-pong, the very nearly optimal rule is to switch at 20.5 total points. This can be seen using the trick of assuming we continue to play to 41 points even if a winner has already been determined. If we switch after point 20, the distribution of scores after point 40 are symmetric for equal players. Hence, the scores $(20 + x, 20 - x)$ and $(20 - x, 20 + x)$ are equally likely. When $x > 0$, the first case implies a win for player 1 while the second implies a win for player 2. The only place for imbalance is the resolution of a 20–20 tie. Let the chance of a 20–20 match be $2p$ so that each player has a $50 - p$ chance of already winning. Since the player with the good side wins the next point with probability q, he wins the game with chance $50 + (2q - 1)p$.

"On the other hand, suppose you switch after point 21. Then the scores after 42 points are symmetric. Again everything is balanced except for the resolution of a 21-21 tie. Let the chance of a tie at 21-21 be $2\hat{p}$. Then the second-half good-side player is at a disadvantage in the 21-21 game, since it is more likely than not that he won the previous point. Reasoning as before, he will have a probability of $50 + (1 - 2q)\hat{p}$ of winning the game. By choosing 20 or 21 point as the switch number via a coin flip, we achieve $50 + (2q - 1)(p - \hat{p})$, which would be very close to 50 percent.

"This also suggests that the winning by two point rule may actually simplify rather than complicate the optimal switching rule. Switch after 20 points. Then if the score is tied at 21-21, switch after the next point and switch again after every second point thereafter until one side is ahead by two."

Keeler continues with a comment on optimal scoring rules. He describes the work of Carl Morris, a statistician at University of Texas. "In most cases, to maximize the chance of the better player winning, while constraining the expected number of points played, you should play until one side gets N points ahead. If the server in squash has an advantage, then it is more informative to change the rule so that the person who loses the point gets the next serve, and you only win points when you are the receiver." Since they have changed the ball in squash 3 times in the last 15 years, maybe they will catch on and change the rules.

■ *Kaushik Basu, Avinash Dixit, Jack Hirshleifer, Emmett Keeler, Eliot Robin, Tom Romer, and Richard Zeckhauser share the credit for this issue—my heartfelt thanks. Behind-the-scenes, comments and kibitzing from Timothy Taylor and Carl Shapiro ensure that each puzzle isn't unnecessarily puzzling.*

References

Conniff, Richard, "Superchicken: Whose Life is it Anyway?" *Discover*, June 1988, 33-41.

Grossman, Sanford J., and Oliver D. Hart, "Takeover Bids, the Free-Rider Problem, and the Theory of the Corporation," *Bell Journal of Economics*, Spring 1980, *11*, 42-64.

Poirier, Dale J., "Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics," *Journal of Economic Perspectives*, Winter 1988, *2*, 121-144.

Journal of Economic Perspectives — Volume 2, Number 4 — Fall 1988 — Pages 191–194

Recommendations for Further Reading

Bernard Saffran

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, the articles chosen will be expository or integrative and not focus on original research. As a rule of thumb, articles listed in this section appear in journals that are not indexed by the *Journal of Economic Literature* (which focuses on the research journals) and are similar to many of the articles published in this journal. They may include survey articles, discussion of related subjects in which economists might have an interest (for example, the use of game theory in evolutionary biology) or analyses of economics from other perspectives. The intention is to publish a selective list of ten to fifteen articles per issue, with the selections depending heavily on input from readers. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Bernard Saffran, c/o *Journal of Economic Perspectives*, Department of Economics, Swarthmore College, Swarthmore, Pennsylvania 19081.

Speaking of Art and Prisoners

These two papers coauthored by Bruno S. Frey illustrate how current theory can be applied to issues that would interest undergraduates. They are the sort of papers that could easily stimulate a spirited class discussion.

Frey, Bruno S., and Werner W. Pommerehne, "Is Art Such a Good Investment?" *The Public Interest*, Spring 1986, 79–86. As the form of the question indicates, the

■ *Bernard Saffran is Professor of Economics, Swarthmore College, Swarthmore, PA.*

answer given in this article is "no." The authors have updated a number of studies and show that the real rate of return is below that for public securities: from 1635 to 1987 the average real rate of return on paintings was 1.5 percent, while that on government securities was above 3 percent. This paper is accessible to introductory economics students and uses an interesting topic to illustrate efficient markets and opportunity costs.

Frey, Bruno S., and Heinz Buhofer, "Prisoners and Property Rights," *The Journal of Law and Economics*, April 1988, 19-46. A history of the treatment of military captives organized around the theme of the changing property rights held by the captors over the prisoners.

Amalgamated Articles

These articles range from the economics of teaching to the economics of the family, from measurements by economists to measurements by statisticians, and all the way to explaining poverty in the 1980s.

Becker, Gary C., and Kevin M. Murphy, "The Family and the State," *The Journal of Law and Economics*, April 1988, 1-18. Delivered as the Henry Simons lecture to the University of Chicago Law School. They "consider not only subsidies to education and training but also social security and other old-age support, subsidies to births, laws that limit access to divorce and the sale of children, and laws that require parents' permission for early marriage and other choices of children." The authors conclude, "It is remarkable how many state interventions in family decisions appear to contribute to the efficiency of family arrangements."

Bergmann, Barbara R., "'Measurement' or Finding Things Out in Economics," *Journal of Economic Education*, Spring 1987, 191-201. A critical discussion of economists' methods. She argues, "Most of what we represent ourselves as knowing we have learned from a 'theory' derived from musing at length on the implications of a few of the facts we know about the economy. The crumbs of factual content in our theorizing don't come from regressions on government data, but have been gathered casually in the course of our everyday lives."

Rosen, Sherwin, "Some Economics of Teaching," *Journal of Labor Economics*, October 1987, 561-575. A far ranging essay on the institutions of the higher education industry, especially methods of payment for teachers, that utilizes industrial organization theory like transaction cost and agency theory. Rosen, along the way, analyzes Adam Smith's proposal that "tied a teacher's income to teaching quality as assessed by students," claims that the implicit student fee is $4.00 per lecture for college teachers as compared to $1.30 per class meeting in primary and secondary schools, and argues "that the most important constraint in the economics of the education industry is the student-teacher ratio."

Focus, "Welfare Reform and Povery" Spring 1988, published by the Institute for *Research on Poverty, University of Wisconsin-Madison. This special issue has a* discussion of the nature of the new poverty of the 1980s as well as suggestions for its amelioration by a group of social scientists, including Daniel Moynihan and Michael Novak, all of whom have been concerned with these problems for many years.

As previous columns have indicated, there appears to be considerable debate among statisticians as to the appropriate applications of various statistical models. The debate continues with Hodges, James S., "Uncertainty, Policy Analysis and Statistics," and commentators, *Statistical Science*, August 1987, 259–291. Hodges is a statistician with the RAND corporation. Two of the commentaries by David Freedman and Albert Mandansky are of particular interest.

Targets of the Fed

As always, the publications of the Federal Reserve banks include some useful and interesting articles.

Strongin, Steven, and Paula S. Binkley, "A Policymakers' Guide to Economic Forecasts." *Economic Perspectives*, Federal Reserve Bank of Chicago, May/June 1988, 3–10. The moral of the article is a pessimistic one. "This study analyzes the accuracy of forecast revisions... Forecast revisions, *during the year being forecast*, are as likely to worsen as to improve the accuracy of the forecast with as much as six months' additional information... Early forecasts were often more accurate than their mid-term forecast revisions."

The *Business Review* of the Federal Reserve Bank of Philadelphia continues to be a major source of articles that are helpful to teachers and accessible to undergraduates. Two such articles focusing on new developments on how markets function appeared in the March/April 1988 issue. They are:

Mester, Loretta J., "Going, Going, Gone: Setting Prices With Auctions," 3–13.

Taylor, Herb, "Experimental Economics: Putting Markets Under the Microscope." 15–25.

People Who Are Also Personalities

O'Brien, D. P., "Lionel Charles Robbins, 1898–1984," *The Economic Journal*, March 1988, 104–125. "Lionel Robbins was a leading figure in what Harry Johnson has called 'the second golden age of British economics', a leading British academic, a prolific writer on economics, and a majestic public figure whose fame extended far beyond the narrow academic community."

For those who followed the debate between Simon and Koblitz that was discussed in my previous column, there are rejoinders and replies in *The Mathematical Intelligencer*, Spring 1988, 10–12.

Matinee Request

I have always been somewhat envious of my colleagues in other fields who, on appropriate occasions, show a movie. I would like to put together a column, a sort of "Recommendations for Further Viewing," that would list movies suitable for use in the classroom. For example, Satyajit Ray's "Distant Thunder" is a visual complement to A. K. Sen's *Poverty and Famines* while Frank Capra's "It's a Wonderful Life" would serve as the basis for a discussion of reputation effects and the self-fulfilling prophecy in discussing bank runs. I invite readers to send suggestions with as much detail as you have available.

■ *I would like to thank Charles Lave for his suggestions.*

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 195–206

Anomalies
The Ultimatum Game

Richard H. Thaler

Economics can be distinguished from other social sciences by the belief that most (all?) behavior can be explained by assuming that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear. An empirical result qualifies as an anomaly if it is difficult to "rationalize," or if implausible assumptions are necessary to explain it within the paradigm. This column will present a series of such anomalies. Readers are invited to suggest topics for future columns by sending a note with some references to (or better yet copies of) the relevant research. Comments on anomalies printed here are also welcome. The address is: Richard Thaler, c/o *Journal of Economic Perspectives*, Johnson Graduate School of Management, Malott Hall, Cornell University, Ithaca, NY 14853.

Introduction

Imagine yourself in the following situation. Your daughter Eve, off at college, calls you to ask for your sage advice. She has agreed to participate in a laboratory experiment being run in the economics department at her college. The rules were explained in advance so that the subjects could think carefully about their choices. The experiment involves two-player bargaining, with Eve placed in the role of Player 1. She is to be given $10, and will be asked to divide it between herself and another student (Player 2) whose identity is unknown to her. The rules stipulate that she must make Player 2 an offer, and then Player 2 can either accept the offer, in which case he will receive whatever Eve offered him, or he can reject the offer, in which case both

■ *Richard H. Thaler is Henrietta Johnson Louis Professor of Economics, Johnson School of Management, Cornell University, Ithaca, New York.*

players will receive nothing. Her question to her wise economist parent: How much should she offer?

Before answering, you decide to check the relevant theory, in this case a paper by Rubinstein (1982) (see also Stahl, 1972). You immedately notice that Rubinstein starts his article with the disclaimer that he is only theorizing about what will happen in a bargaining situation if both parties behave rationally. He explicitly distinguishes this question from two others (p. 97) namely: "(i) the positive question—what is the agreement reached in practice; (ii) the normative question—what is the just agreement."

After reading Rubinstein, including his opening disclaimer, you realize that the theory for the simple game Eve has to play is rather obvious. Player 1 should offer Player 2 a penny. Player 2 will accept this offer, since a penny is better than nothing. However, you now realize why Rubinstein was so careful. Offering only a penny seems to be a risky strategy. If Player 2 views such a small offer as insulting, it would cost him only a penny to reject it. Maybe Eve should offer more than a penny? But how much more? What advice would you give?

While mulling over what to tell your daughter, you get a phone call from a local merchant offering you a consulting job, an event about as frequent as your daughter asking for your advice. The merchant owns a local motel in the college town in which you reside. He is troubled by the fact that a few times a year, such as graduation and homecoming weekends, there is enormous excess demand for rooms. On graduation weekend, for example, some parents stay in hotels as much as 50 miles away. The usual price for a room in his motel is $65 a night. Normal practice in town is to retain the usual rates, but to insist on a three night minimum stay. He estimates that he could easily fill the motel for graduation weekend at a rate of $150 a night, while retaining the three night minimum stay. However, he is a bit uneasy about doing this. He is worried about being labeled a "gouger," and thinks this label might hurt his regular business. "You are an economist," he says. "What should I do?" While thinking over this problem you realize that it has something in common with Eve's dilemma, and that you may need more than economic theory to advise either of your new clients. But what?

Simple Ultimatum Games

The game described by Eve is known as an Ultimatum Game. The first experiments to use this game were conducted by three German economists, Güth, Schmittberger, and Schwarze (1982), or GSS. They divided their sample of 42 economics students in half. One group was designated to take the role of Player 1, the Allocator; the other group took the role of Player 2, the Recipient. Each Allocator was asked to divide c German marks (DM) between himself and the Recipient. If the offer x was accepted then the Allocator received $c - x$ and the Recipient received x. If the offer was rejected, both players received nothing. The size of the stake to be divided, c, was varied between DM4 and DM10. Then, a week later, the same subjects were invited to play the game again.

If the Rubinstein model is a good positive model (in spite of his disclaimer) then two results should be observed.[1] First, allocators should make offers approaching zero. Second, recipients should accept all positive offers. The data are inconsistent with both of these predictions. In the first experiment (with inexperienced subjects) the modal offer was a 50 percent split (7 of 21 cases). The mean offer was .37*c*. Two students did ask for all of *c* in games where c = DM4, with one of these offers being accepted,[2] the other rejected. All other offers were for at least DM1, and one positive offer of DM1.20 was rejected.

In the replication, after a week to think about it, the offers were somewhat less generous, but still considerably greater than epsilon. The mean offer was .32*c*, and only two players offered an even split. However, there was only one offer of less than DM1 and it was rejected. Also, three offers of DM1 were rejected as was an offer of DM3. Thus 5 of the 21 offers were rejected.

Both the Allocators and the Recipients take actions inconsistent with the theory. The Recipients' actions, however, are easier to interpret. When a Recipient declines a positive offer, he signals that his utility function has non-monetary arguments. The decline of an offer of .1*c* says, "I would rather sacrifice .1*c* than accept what I consider to be an unfair allocation of the stake." The extent of this willingness to decline positive but unfair offers is explored below. The actions of the Allocators could be explained by either of two motives (or some combination of both). Allocators who make significantly positive offers could either have a taste for fairness, and/or could be worried that unfair offers will be (rationally or mistakenly) rejected. Further experiments reveal that both explanations have some validity.

GSS investigated the behavior of Recipients in a second experiment using 37 new subjects. In this study, subjects were told they would play the game twice, once as Allocator and once as Recipient. In all games, c = DM7. They were asked to make an offer as Allocator, and to indicate the minimum payment they would accept when they played the role of Recipient. (Note that these are real contingent responses, not answers to hypothetical questions.) The Allocators' responses in this experiment were even more generous than those observed in the earlier experiments, the mean offer being .45*c*. Of greater interest are the responses of the subjects as Recipients. All but two of the subjects indicated a reservation demand of at least DM1, and the median reservation demand was DM2.50.

Two related experiments were conducted by Kahneman, Knetsch, and Thaler (1986b), or KKT. In the first, conducted at the University of British Columbia, the GSS study was replicated to determine whether the results might be caused by subjects being confused about the task. A simple ultimatum game was played, with c = $10 (Canadian). Again subjects were asked to say what they would do in both roles. Two steps were taken to be sure that the subjects understood the task. First, the subjects were asked two preliminary diagnostic questions. Of the 137 subjects who participated in the study, 22 were dropped because they did not answer both questions correctly. Second, rather than asking subjects to directly state their reservation demand, the subjects were asked a series of yes or no questions of the form: If the other player offers you \$.50, will you accept the offer or reject it?[3] These questions were repeated in increments of 50 cents. In three different experiments, the mean minimum acceptable offer varied between \$2.00 and \$2.59, amounts similar to those obtained by GSS.[4]

---

[1] Actually, Rubinstein has pointed out to me that these predictions are not derived from game theory per se, but depend also in additional assumptions, such as that the players are expected utility maximizers with path independent utility functions (i.e., they get utility from the outcome of the negotiation, not the process).

[2] We can't be sure whether the Recipient who accepted the zero offer was confused, generous, or simply had a deep understanding of bargaining theory.

The second KKT experiment investigated two questions. First, will Allocators be fair even if their offers cannot be rejected, and second, will subjects sacrifice money to punish an Allocator who behaved unfairly *to someone else*. In the first part students in a psychology class at Cornell University were asked to divide \$20 between themselves and another anonymous member of the class. They were given only two choices of allocations: they could keep \$18 for themselves and give their partner \$2, or they could offer an even split of \$10 each. (At these stakes it was not possible to have a large sample size and still pay everyone. Thus, the subjects were told that eight pairs of students would be selected at random and paid.) Unlike the previous experiments, the offers made by the Allocators could not be rejected by the Recipients. Nonetheless, offers were still very generous. Of the 161 subjects, 122 (76 percent) divided the \$20 evenly. Therefore, part of the explanation for the generous offers observed in the ultimatum game does appear to be explained by a taste for fairness on the part of the Allocators.

After completing the first part of the study, the same students were given another question. They were told they would be matched with two students who had not been selected to be paid in the first part of the experiment. One of these students had taken the \$18 (called U for uneven) while the other had taken \$10 (E). A subject was then asked to choose between the following: He could take \$6 for himself and give \$6 to U, or he could take \$5 for himself and give \$5 to E. Thus the question came down to whether subjects would be willing to pay a dollar to split money with a stranger who had been generous rather than split with a stranger who had been greedy. A clear majority, 74 percent, elected to take the smaller reward in order to split with E.

Two-stage Bargaining Games

GSS (1982, p. 385) conclude that game theory is "of little help in explaining ultimatum bargaining behaviour." With the honor of game thory at stake (or at least its descriptive validity) game theorists Binmore, Shaked, and Sutton (1985), BSS, performed a pair of experiments. They revised the GSS design by adding a second stage to the bargaining game and had the players communicate via linked microcomputers. The two-stage game begins as before with Player 1 in the role of Allocator, Player 2 in the role of Recipient, and with $c = 100$ UK pence. The allocator makes an offer of x (keeping $c - x$ for himself). If this offer is refused, then the game moves to round 2, with the players reversing roles and the stake reduced to δc, where the discount factor δ in this case was set at 0.25. The second round is a simple ultimatum game with $c = 25$p and Player 2 in the role of Allocator. The (subgame perfect) equilibrium for this game is found through a trivial backward induction. If the game reaches round 2, then Player 2 can offer Player 1 just a penny, retaining 24p for himself. Therefore, Player 2 will accept anything more than 24 in round one, so Player 1 should offer 25p on round one.

---

[3] In other research, Knetsch, Thaler, and Kahneman (1988) have found that this sort of yes or no question is easier for the subject to answer. For example, subjects are more likely to give the correct (demand revealing) answers in a second price auction when the questions are posed in this way.

[4] The three experiments had different groups of students as subjects. In all cases they were told that their partner would be someone in another class. The offers of the Allocators were similar to those obtained by Güth et al., with the mean amount offered ranging from \$4.21 to \$4.76. Of interest is the fact that the most generous offers were made by students in a psychology class making offers to students in another psychology class. The psychology students were less generous when making offers to students in a commerce class, but the least generous offers were made by commerce students to the psychology students. Similarly, the commerce students indicated the smallest minimum acceptable offer.

This game was played twice. In the first game, offers by the Allocators were similar to those observed in earlier experiments. The modal offer was 50 pence, and only 10 percent were in the range 24–26 pence. Also, 15 percent of the first-round offers were rejected (whereas the theory predicts the game will never reach the second round). In the second game, the subjects who had played in the role of Player 2 in the first game were invited to play another game, this time in the role of Player 1. (Responses of their hypothetical partners were not collected.) This time the subjects behaved more in accordance with game theory. The modal offer was just below the equilibrium of 25p. The authors conclude (p. 1180) that considerations of fairness "are easily displaced by calculations of strategic advantage, once players fully appreciate the structure of the game." However, three aspects of the BSS experiments raise questions about how to interpret their results.

First, the subjects were not informed of the existence of a second round until after the first round was played. If subjects thought that the game was now one where they would take turns being Player 1, they may have felt that alternatively taking the equilibrium $.75c$ would average out to a fair distribution.

Second, in conducting their experiments, BSS took the unusual step of telling their subjects how to behave. Specifically, the written instructions included the following passage: "*How do we want you to play?* YOU WILL BE DOING US A FAVOUR IF YOU SIMPLY SET OUT TO MAXIMIZE YOUR WINNINGS." (Emphasis and all caps in the original.) It is difficult to say what effect such instructions might have on the results without a controlled experiment (though it is reassuring that the first round results are similar to those obtained by GSS). However, in another similar context instructions did prove to have a powerful effect. Hoffman and Spitzer (1982) ran an experiment which is very similar to the ultimatum game. The Allocator (who was given that role as a result of a coin flip) could choose between an outcome which gave him $12 and the recipient nothing, or, if both players agreed, they could divide $14. Of course, the theory predicts that the players will agree to divide the $14, with the Allocator getting no less than $12. Instead, all pairs agreed to split the $14 evenly, getting $7 each. In a second paper Hoffman and Spitzer (1985) tried to understand why this happened. Two manipulations were crossed with each other to produce four conditions. In the first, the role of Allocator was determined either by a coin flip, or by playing a simple game with the winner becoming the Allocator. In the second, winners of the coin flip or game were told either that they had "earned" the right to be the Allocator, or that they were "designated" as Allocator. Of the two manipulations, the second was the more powerful. The difference between the game and the coin flip was not significant, but the subjects who had been told that they had "earned" the property right took significantly more of the money. Further research on this type of demand characteristic is clearly needed.

Third, the two-stage game devised by BSS differs from the simple ultimatum game in one key respect. The equilibrium offer of 25p is distinctly positive. This means that compared to the simple ultimatum game, it is more costly for a Recipient to reject the equilibrium offer, and the equilibrium offer is more fair. To see whether these factors are important, Güth and Tietz (1987) tried a two-stage game with a discount factor of .1 or .9. When $\delta = .1$ the equilibrium offer is a rather unfair $.10c$, while when $\delta = .9$ the equilibrium offer is a full $.90c$ (hardly fair to oneself!). The games were played twice with players switching roles.[5] The stake was either DM5, DM15, or DM35.

The results of these experiments did not support the BSS conclusion that rationality will take over if the players have a chance to think about the game. In the trials where $\delta = .1$, offers increased (moved away from equilibrium) from trial one to trial two (from $.24c$ to $.33c$). For the cases where $\delta = .9$, the mean offers also increased on trial two (from $.37c$ to $.49c$), which is toward the equilibrium value. Averaging across both trials and all levels of c, the mean offers when $\delta = .1$ were $.28c$, while when $\delta = .9$ the mean offers were $.43c$. Neither is close to their respective equilibrium values of $.1c$ and $.9c$. The variation in the level of c also provides some evidence on the robustness of the phenomena under study. If we compare the games played with $c =$ DM5 to those with $c =$ DM35, we find that the offers move part way toward the equilibrium levels (from $.33c$ to $.24c$) when $\delta = .1$, and slightly away from equilibrium (from $.36c$ to $.34c$) to $\delta = .9$. Thus, raising the stakes does little to improve the descriptive value of game theory.[6]

Multi-stage Games

The next contribution to the analysis of ultimatum games is Neelin, Sonnenschein, and Spiegel (1987) (NSS). Subjects in their experiments were Princeton undergraduates enrolled in an intermediate microeconomics class. Subjects played a series of games with the number of periods (announced in advance) varying between 2 and 5, and $c = \$5$. Player 1 makes an offer in odd-numbered rounds, and Player 2 in even-numbered rounds. If the final round offer is rejected then both players get nothing. The discount rates were varied in such a way that the equilibrium offer in the first period was always $\$1.25 + \varepsilon$ (or $1.26). In the two-period game the second period c is $1.25; in the three-period game c falls first to $2.50 and then to $1.25; in the five-period game the values for c are $5.00, $1.70, $ .58, $ .20, and $ .07.[7] Subjects first played a practice (4-round) game then played the 2, 3, and 5-round games in that order, each with a different anonymous partner. Subjects retained the same role in each game.

---

[5] One additional rule was put in place. Player 2 could not reject an offer and respond with a counteroffer that gave himself less than he had been offered. Such actions constituted disagreement with both players receiving zero. Thus when $\delta = .1$ if Player 1 offered more than $.1c$ this amounted to an ultimatum since if Player 2 rejected the offer disagreement was declared. The experiments by Ochs and Roth (1988) discussed below show that this rule was probably binding.

[6] What would happen in an ultimatum game with $c = \$1000$, or $\$100,000$? None of us have the research funds to run this experiment, so we can only guess. My own guess is that Recipients' minimum acceptable offers would increase with c, but not linearly. When $c = \$10$, the median minimum acceptable offer is about $.2c$. For $c = \$1000$ I would guess it would fall in the range $.05c$–$.1c$ (\$50–100). The minimum acceptable offer probably also increases with wealth, implying that resisting unfair offers is a normal good.

The idea behind the NSS design is that the results of the various length games can be compared to avoid conclusions that are special to a particular game. The value of the design is quickly appreciated when the results are examined. In the two-round games, the game theoretic prediction did pretty well. Of the 50 Allocators (whom NSS call "sellers"), 33 made offers between $1.25 and $1.50 (the equilibrium value is $1.26). These results are similar to those obtained in the second BSS experiment. In the three-round game, however, the results are completely different. Here 28 out of the 50 players offered an even split of $2.50, with nine others making offers within $ .50 of this amount. Remember that the equilibrium offer in this game is still $1.26.

The five-round game yielded yet another pattern of results. The modal (14) first-round offer was $1.70 and 33 of the 50 offers were in the range $1.50-2.00. NSS note that the players seem to have adopted the strategy of offering Player 2 the stake to be played for in round 2. This *is* the equilibrium offer in the two-stage game, but not in the longer games. Such a strategy might be adopted if players are myopic, and only think one step ahead, or are just conservative, wishing to minimize the risk that their partner will reject their offer for rational or irrational reasons.

NSS conducted a second experiment in which subjects played the five-round game four times with all the payoffs increased by a factor of 3 ($c = \$15$). The results were essentially unchanged. Seventy percent of the offers were in the range $5.00–5.10 (the second round stake is $5.10). No offer close to the equilibrium $3.76 was observed. There was also no evidence of any learning. That is, there was no apparent trend in the offers over the four trials.

By far the most ambitious set of experiments conducted to date is reported in Ochs and Roth (1988). They introduced the following innovations. First, subjects complete 10 bargains, one after another, with all parameters held constant (but with a different opponent each time).[8] This feature allows for a test of whether subjects learn to be proper economists with practice. Second, discount rates were varied separately for each subject. This was accomplished by having subjects bargain for 100 "chips."

---

[7] Notice that the backward induction necessary to derive the equilibrium first round offer is a bit more complicated in the three-and five-round games. The analysis for the five-round game is if the game reaches the fifth round, Player 1 is the Allocator, and he can offer Player 2 a penny (which Player 2 will, by assumption, accept) so Player 1 can get 6 cents at this stage. This implies that at the fourth stage, Player 2 must offer Player 1 at least 6 cents, keeping 14 cents for himself, and so forth.

[8] Subjects were told that at the completion of the experiment that one of the rounds would be selected at random, and they would be paid based on their outcome in that round.

In the first round of any game the chips were worth \$ .30 to each player (so $c = \$30$). In the second round the chips would be worth δ_1 (\$ .30) to Player 1 and δ_2(\$ .30) to Player 2. In the third round, for three-round games, the discount rates were squared. The two discount rates were common knowledge, but were not necessarily equal. Four combinations for (δ_1, δ_2) were varied experimentally: (.4, .4), (.6, .4), (.6, .6), and (.4, .6). These four conditions were crossed with the number of periods to be played *(either 2 or 3) to produce a 4 × 2 experimental design.*

The authors use this complicated experimental design to test two implications of bargaining theory. First, Player 1's discount factor should only matter in games with three periods. (Work through the backward induction to see why.) Second, holding the discount rates constant, Player 2 should receive less in three-period games than in two-period games. (This is true because in three-period games Player 1 gets to make both the first and the last offer.) Also, the theory yields predictions of all the 28 pairwise comparisons between the cells of the experiment.

The results of these experiments provide little support for the descriptive value of game theory, even on the last trials of the experiments. The theory performed well in only one of the eight cells. In the other seven cells, the theoretical mean offer was never within two standard deviations of the actual mean on any trial. Also, both of the additional predictions mentioned in the previous paragraph failed. The Player 1 discount rate mattered in games when it shouldn't, and the length of the game didn't matter when it should. As one simple measure of the ability of the theory to explain the data, Ochs and Roth regressed the observed mean offer on the theoretical offer for the last trials of each cell of the experiment. The R^2 for this equation was .065, and the coefficient on the theoretical offer less than one standard deviation away from zero.

Ochs and Roth also replicate the earlier findings by GSS and KKT regarding Recipients' willingness to decline positive but unfair offers. In these games, if players cared only about monetary payoffs, then Player 2 would never reject Player 1's initial offer and subsequently demand less for himself in his counter offer. Yet Ochs and Roth find that for 81 per cent of the counterproposals, Player 2 demands less cash than he was originally offered by Player 1. The conclusion that subjects' utility functions have arguments other than money is reconfirmed.

We have seen that game theory is unsatisfactory as a positive model of behavior. It is also lacking as a prescriptive tool. While none of the subjects in Ochs and Roth's experiments came very close to using the game-theoretic strategies, those who most closely approximated this strategy did not make the most money. In fact, in 4 of the 8 cells, the player with the highest average demand (over the ten trials) had the *lowest* average earnings.

Ultimatums in the Market

The willingness of people to resist what they consider to be unfair allocations has implications for economics that go well beyond bargaining theory. Any time a monopolist (or monopsonist) sets a price (or wage), it has the quality of an ultimatum. Just as the Recipient in an ultimatum game may reject a small but positive offer, a buyer may refrain from purchasing at a price that leaves a small bit of consumer surplus but is viewed as dividing the surplus in an unfair manner. Consider the following problem posed to two groups of participants in an executive education program (Thaler, 1985). One group received a version with the passages in brackets, the other the passages in parentheses.

> You are lying on the beach on a hot day. All you have to drink is ice water. For the last hour you have been thinking about how much you would enjoy a nice cold bottle of your favorite brand of beer. A companion gets up to go make a phone call and offers to bring back a beer from the only nearby place where beer is sold (a fancy resort hotel) [a small run-down grocery store]. He says that the beer might be expensive and so asks how much you are willing to pay for the beer. He says that he will buy the beer if it costs as much or less than the price you state. But if it costs more than the price you state he will not buy it. You trust your friend, and there is no possibility of bargaining with the (bartender) [store owner]. What price do you tell him?

Notice that the scenario here is a simple ultimatum game with the respondent in the role of the Recipient. The median response for the fancy hotel version was $2.65, while the median for the grocery store version was $1.50. Because of a difference in perceived costs, the price of $2.65 seems fair for a resort hotel, but a "rip-off" in a run-down grocery store.

In general, consumers may be unwilling to participate in an exchange in which the other party gets too large a share of the surplus. This may explain why some markets (Super Bowl tickets, reservations at the most popular restaurant in town on Saturday night, Bruce Springsteen concert tickets) fail to clear at the official price set by the seller. Whenever the seller has an ongoing relationship with the buyer and the market clearing price would be considered unfairly high, the seller has an incentive to keep prices below the equilibrium in order to retain future business. (These issues are discussed in more detail in Thaler, 1985, and Kahneman, Knetsch and Thaler, 1986a.)

Commentary

Bell, Raiffa, and Tversky (1988) have suggested that it is useful to distinguish three kinds of theories of decision making under uncertainty. *Normative* theories tell us how a rational agent should behave. *Descriptive* theories tell us how agents do behave. *Prescriptive* theories offer advice as to how to behave when faced with one's own cognitive or other limitations. The research on the bargaining games indicates that we need a similar triple of game theories. Game theory as it currently exists is a normative theory. It characterizes optimal behavior when selfishness and rationality are common knowledge. Experimental research is starting to provide the evidence necessary to formulate a good description of how people actually behave. However, as yet we have little research that would help develop prescriptive game theory. The analysis of Eve's problem illustrates this gap in our repertoire. To solve for the income maximizing offer, one would have to be able to characterize the acceptance function for the recipient. For any given offer, what is the probability it will be rejected by the recipient?

In multi-stage games, the optimal strategy is even less clear. Consider the five-stage game in NSS where $c = \$15$. The values for c in the second through fifth stages of the game are: \$5.10, 1.74, .60, and .21. What is an optimal offer at stage 1? There are two important prescriptive game theoretic considerations. What offer will Player 2 consider fair? Does Player 2 understand the game? Both factors may be important. To get a sense of the possible role of the second factor, I arranged to have a question posed on the final exam for an MBA level course on "Pricing and Strategy" at Cornell. The course has intermediate microeconomics as a prerequisite, and the students had discussed game theory, backward induction, and simple ultimatum games in class. The exam consisted of eight questions from which the students had to answer five. The question of interest began with a description of the five-round game played in NSS. The students were told to assume that both players are rational, and both wish to maximize the money they earn in this game. They were then asked: What is the smallest offer Player 1 can make in round 1 which be accepted by Player 2?

Of the 30 students in the class, only 13 chose to answer this question, and only 9 answered it correctly. This implies that more than half the class was not sure they knew the answer to the question, and of those who did think they knew the answer, 30 percent got it wrong. Clearly, this is not a trivial question, and backward induction is not an intuitively obvious concept. To see the importance of this issue, consider a Player 1 who is thinking about making an offer of \$4.00 to Player 2. While Player 1 may know that this is more than Player 2 can hope to get if he rejects the offer, if Player 2 thinks he can get \$5.09, he may mistakenly turn the offer down.

So, if Eve were playing this five-round game, before giving her any advice we would want to know how smart her opponent is. Has he studied game theory? Does it look as if he can subtract, much less perform a backward induction? More generally, in order to develop prescriptive game theory, the assumption that rationality and wealth maximization are common knowledge will have to be modified. A rational wealth maximizing player must realize that his opponent may be neither, and make appropriate changes to his policies.[9] Notice that in developing prescriptive game theory, it is necessary to do both theory and empirical work. Theory alone cannot tell us what factors enter our opponent's utility function, nor what bounds must be placed on his rationality.

---

[9] This sort of analysis is common in expert bridge. In tournament bridge, experts often play against non-experts, unlike many other competitive events. Optimal strategy in a weak field depends in part on giving one's opponents numerous opportunities to make mistakes.

One conclusion which emerges clearly from this research is that notions of fairness can play a significant role in determining the outcomes of negotiations. However, a concern for fairness[10] does not preclude other factors, even greed, from affecting behavior. In their article, BSS pose the problem starkly as a contest between two extreme positions. People are thought either to be "fairmen" who divide everything equally, or "gamesmen" who behave selfishly and rationally like proper economic agents. I think it is safe to say that most people are not well described by either extreme view. Rather, most people prefer more money to less, like to be treated fairly, and like to treat others fairly. To the extent that these objectives are contradictory, subjects make trade-offs.[11] Behavior also appears to depend greatly on context and other subtle features of the environment. In some experiments most Allocators choose even splits, in others most choose the game-theoretic allocation. Future research should investigate the factors that produce each kind of behavior, rather than attempt to demonstrate that one type of behavior or the other predominates.

Just as the characterization of the behavior of subjects as either fairmen or gamesmen is too simplistic, so is any distinction on a "hard" vs. "soft" dimension. There is a tendency among economists to think of themselves, and the agents in their models, as having hard hearts (as well as heads, noses, and other extremities). *Homo economicus* is usually assumed to care about wealth more than such issues as fairness and justice. In contrast, many economists think of other social scientists (and the agents in *their* models) as "softies." The research on ultimatum games belies such easy characterizations. There is a "soft" tendency among the allocators to choose 50–50 allocations, even when the risk of rejection is eliminated. Yet the behavior of the recipients, while inconsistent with economic models, is remarkably hard-nosed. They say, in effect, "Take your offer of epsilon and shove it!"

■ *I wish to thank Ken Binmore, Julia Grant, Daniel Kahneman, Ariel Rubinstein, Carl Shapiro, and Hal Varian for helpful comments.*

---

[10] It must be emphasized that issues of fairness are complicated. Perceptions of fairness often diverge from those which seem natural to economists. For example Kahneman, Knetsch, and Thaler (1986a) found that most people believe a queue is more fair than a market, and Yaari and Bar-Hillel (1984) found that when making judgments of justice, people distinguish between "needs" and "wants." Fairness *arguments* are also quite common in negotiations. While bargainers use fairness arguments for self-serving reasons ("I think I should get more because that would be fair . . .") such arguments can nonetheless be effective (Roth, 1987).

[11] To illustrate, in the experiment conducted by Kahneman et al. (1986b) Allocators were permitted to choose between just two divisions of $20, either $18–2 or $10–10. Most chose the even allocation. However, had they been allowed to choose an intermediate allocation, such as $12–8, many might have selected that.

References

Bell, David, Howard Raiffa, and Amos Tversky, "Descriptive, Normative, and Prescriptive Interactions in Decision Making." In Bell, David, Howard Raiffa, and Amos Tversky, eds. *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge, England: Cambridge University Press, 1988.

Binmore, Ken, Avner Shaked, and John Sutton, "Testing Noncooperative Bargaining Theory: A Preliminary Study, "*American Economic Review*, 1985, *75*, 1178–80.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 1982, *3*, 367–88.

Güth, Werner, and Reinhard Tietz, "Ultimatum Bargaining for a Shrinking Cake: An Experimental Analysis," J. W. Goethe-Universität, unpublished, 1987.

Hoffman, Elizabeth, and Matthew L. Spitzer, "The Coase Theorem: Some Experimental Tests," *Journal of Law and Economics*, 1982, *25*, 73–98.

Hoffman, Elizabeth, and Matthew L. Spitzer, "Entitlements, Rights and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice," *Journal of Legal Studies*, 1985, *14*, 259–97.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, 1986a, *76*, 728–41.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Fairness and the Assumptions of Economics," *Journal of Business*, 1986b, *59*, S285–S300.

Knetsch, Jack L., Richard H. Thaler, and Daniel Kahneman, "Experimental Tests of the Endowment Effect and the Coase Theorem," unpublished manuscript, Cornell University, Johnson Graduate School of Management, 1988.

Neelin, Janet, Hugo Sonnenschein, and Matthew Spiegel, "A Further Test of Bargaining Theory," unpublished, Princeton University, Department of Economics, 1987.

Ochs, Jack, and Alvin E. Roth, "An Experimental Study of Sequential Bargaining," unpublished, University of Pittsburgh, Department of Economics, 1988.

Roth, Alvin E. "Bargaining Phenomena and Bargaining Theory." In A. E. Roth ed., *Laboratory Experimentation in Economics: Six Points of View*. Cambridge, England: Cambridge University Press, 1987.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 1982, *50*, 97–109.

Stahl, Ingolf, *Bargaining Theory*, Economic Research Institute, Stockholm, 1972.

Thaler, Richard, "Mental Accounting and Consumer Choice," *Marketing Science*, Summer 1985, *4*, 199–214.

Yaari, M., and Maya Bar-Hillel, "On Dividing Justly," *Social Choice and Welfare*, 1984, *1*, 1–24.

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 207–212

Correspondence

To be considered for publication in the Correspondence section, letters should be relatively short—generally less than 1000 *words—and should be sent to the journal offices at the address appearing inside the front cover. The editors will choose which letters will be published. All published letters will be subject to editing for style and length.*

Comment on "Is Free Trade Passé?" by Paul R. Krugman, Fall 1987, pp. 131–145.

I would like to register a disappointment with the conclusion of an otherwise informative essay by Paul R. Krugman. After outlining recent advances in international trade theory, which incorporate notions of imperfect competition and scale economies to show that free trade can lead to suboptimal outcomes, Krugman retreats into the comfort of the conventional wisdom. He does not defend laissez-faire by arguing that markets are efficient, or by denying that sophisticated intervention could improve on market solutions; instead, he suggests that the potential benefits of such a course are probably small, would be captured by special interest groups, and might have unforeseen political consequences. I would have welcomed a different set of conclusions.

Perhaps one could learn from the historical record, according to which a large number of societies were able to devise trade strategies to their ultimate benefit. Shouldn't we try? For instance, the British achieved commercial hegemony in the early-modern period not by outcompeting the Dutch but by adopting a set of stringent laws that discriminated heavy in favor of their own subjects. Before they adopted free trade, such restrictive legislation served them well for a couple of centuries. Moreover, the American industrialization drive of the antebellum decades did not occur under a regime of free trade. In the late nineteenth century, too, a large number of economies, including those of Germany and the United States, all thrived with protection, and Great Britain, the one country which continued to cling to free trade, found itself outmaneuvered and rapidly fell behind. In our own time, too, Japan did not grow by opening its market to outside competitors. It is difficult to believe that it would have performed even better had it done so. Thus, tariff protection and economic growth are by no means mutually exclusive. No doubt counterexamples could be provided when protection was unsuccessful in fostering growth, but the question is whether there is a lesson to be derived from those instances in which it did.

Perhaps the article might have concluded by suggesting that the advocates of free trade exaggerate, in any event, the applicability of the standard trade theorems to real world situations. These theorems show that free trade increases welfare, but under a very restrictive set of assumptions. More complex models—those which do away with simple barter by introducing money, or in which full employment of resources need not prevail at all times, or in which there are many goods traded between many countries, or which are dynamic rather than static, or which might incorporate nonlinearities or chaos, or which maximize growth rather than short-term utility, or which allow a nation's capital stock to be traded away—are not so clear in their policy implications as the Ricardian model so ingrained in the subconscious of economists.

One might add that theory does not help in predicting how long the market will take to reach a new equilibrium. Can intervention help speed up the equilibrating process? How confident can we be that these complications are trivial, a concern only for the pedantic? Would it not be prudent to urge caution instead because, in actuality, economists still do not understand very well what makes economies grow or decline in the long run?

Krugman warns us not to enter the uncharted waters of intervention because of possible political spin-off effects. Yet all policy, even one of laissez-faire, have unwanted and unforeseen side effects. France, for instance, opened its market to British goods in 1786, which led to severe dislocations in the industrial sector, contributing thereby to the crisis atmosphere that helped lead to the violent outbreak three years later. Note furthermore that our free trade policy vis-à-vis OPEC has now brought us into military conflict in the Persian Gulf. Had the United States adopted import-substitution strategies to reduce substantially our dependence on foreign oil, military intervention in that area might be a less compelling course of action. However, those who argued against the development of alternate energy sources on the basis of a cost-benefit analysis certainly did not factor into their equations the military and human costs we are now incurring.

It should be possible to devise game-theoretic, theory-conforming strategies which both minimize the risk of retaliation and offer a high probability of improving our economic position. Suppose we mandate that our merchandise account be in balance next year. Would our trading partners retaliate by buying less from us, or would it be in their best interest to buy more? Intervention need not inexorably lead to retaliation. We have not, in fact, retaliated against the Japanese for restricting our access to their markets, nor have we retaliated against the efforts of OPEC to restrict free trade in crude oil.

I believe that it would be much more rewarding to confess the limits of our knowledge, and to admit that the case for free trade has been overblown. While free trade has served us well at times, at other times the economy did quite well without it. Given the theoretical uncertainties, why not adopt an experimental approach to trade policy? If GNP falls in response to a new trade policy, we should be able to reverse our course. Even if the gains of a new policy might be "small" at first, accumulated over generations they might turn out to be important, not simply to our well-being, but perhaps even to our very survival. Thus, it seems to me that one could just as well have concluded Krugman's fine essay with a plea for cautious intervention, and for more work on the subject before we allow ourselves to remain dogmatic on the question of free trade.

John Komlos
Associate Professor of History and Economics
University of Pittsburgh

Comment on "The Making of an Economist," by David C. Colander and Arjo Klamer, Fall 1987, pp. 95–112.

A common public perception of economists is that they disagree on a wide variety of issues. Recent studies have sought to determine whether or not there is a factual basis for this perception.[1]

These studies asked professional economists to indicate their general level of agreement with each of thirty economic statements. The authors found that economists do in fact agree on fundamental theory and there is an "economic way of thinking" about important issues which provides a common ground to the approach of economists. This note examines the question of where this agreement begins.

A recent study in this journal examined the opinions of students in six graduate economics programs and found that graduate students generally agreed with professional economists. We were interested in whether undergraduates who have studied economics respond to economic propositions in ways that are similar to professional economists. This may suggest that the economic way of thinking takes root in the first economics courses economists take.

We administered a questionnaire which was virtually identical to that used in previous studies to a group of students who had completed both introductory macro and microeconomics and compared their responses with a group of students who had not taken any college economics. The experimental group consisted of all students completing the introductory macro-micro sequence at the University of Alaska, Anchorage, during the spring semester of 1986. A total of 91 completed responses were

---

[1] Kearl, James R., et al., "A Confusion of Economists?" *American Economic Review Proceedings*, May 1979, 69, p. 2837 and Frey, Bruno S., et. al., "Consensus and Dissension Among Economists: An Empirical Inquiry," *American Economic Review*, December 1984, *74*, pp. 986–994.

Table 1
Proposition Categories and Comparative Responses (Agreement Shown in Percent)

Micro Positive Propositions	*Economists*	*Econ Students*	*Non-Econ Students*
1. Tariffs and import quotas reduce general economic welfare.	97	74*	45*
4. Cash payments are superior to transfers-in-kind.	92	39*	51*
5. Flexible exchange rates offer an effective international monetary arrangement.	95	90	82*
6. A minimum wage increases unemployment among young and unskilled workers.	90	84	31*
15. A ceiling on rents reduces the quantity and quality of housing available.	98	93	62*
17. Effluent taxes represent a better approach to pollution control than imposition of pollution ceilings.	81	83	58*
Micro Normative Propositions:			
11. Antitrust laws should be used vigorously to reduce monopoly power from its current level.	85	59*	76
20. Reducing the regulatory power of federal commissions would improve the efficiency of the U.S. economy.	78	70	68
25. "Consumer protection" laws generally reduce economic efficiency.	52	73*	34
26. The economic power of labor unions should be significantly curtailed.	70	74	73
Macro Positive Propositions:			
3. The money supply is a more important target than interest rates for monetary policy.	71	79*	58*
8. Fiscal policy has a significant stimulative impact on a less than fully employed economy.	92	92	77*
2. Inflation is primarily a monetary phenomenon.	57	64*	44
9. The Fed has the capacity to achieve a constant rate of growth of the money supply if it is so desired.	76	85*	44
4. In the short run, unemployment can be reduced by increasing the rate of inflation.	64	63	29*

obtained from this group. A control group consisting of 68 students who had enrolled in introductory liberal arts courses and had taken no college economics was used for comparison.

Results of the survey are summarized in Table 1. Propositions have been grouped according to micro-positive, micro-normative, macro-positive, and macro-normative categories identified by Kearl, et al. Overall, we found that students who had taken

Table 1, continued

Proposition Categories and Comparative Responses (Agreement Shown in Percent)

Macro Normative Propositions:	*Economists*	*Econ Students*	*Non-Econ Students*
2. The government should be an employer of last resort and initiate a guaranteed program.	53	31*	46
7. The government should index the income-tax structure for inflation.	68	75	78
13. The government should restructure the welfare system along lines of a "negative income tax."	92	58*	49*
16. The Fed should be instructed to increase the money supply at a fixed rate.	39	51*	41*
21. The federal budget should be balanced over the business cycle rather than yearly.	83	45*	69
Miscellaneous:			
9. The distribution of income in the United States should be more equal.	71	43*	59
10. National defense expenditures should be reduced from the present level.	66	55	72*
14. Wage-price controls should be used to control inflation.	28	14	57*
18. The level of government spending should be reduced (disregarding expenditures for stabilization).	57	76*	81*
22. The fundamental cause of the rise in oil prices in the mid-1970's was the monopoly power of the large oil companies.	25	69	78
23. The redistribution of income is a legitimate role for government in the context of the U.S. economy.	81	36*	59*

*Different from Professional Economists at .05 level of significance.

the two-semester, macro-micro sequence agreed with professional economists on 12 of the 26 propositions.[2] Their fellow students who had taken no economics agreed with professional economists on 10 propositions.

The area of greatest agreement between economics students and professional economists was micro-positive propositions where economics students agreed with economists on four of six items, including flexible exchange rates (5), minimum wage laws (6), ceilings on rents (15), and effluent taxes (17). In contrast, students who had taken no economics failed to agree with professional economists on any micro-positive proposition. They did agree with professionals on four micro propositions, but all of these were classified as normative.

---

[2] Agreement between groups was determined by the significance of the difference in their answers. As in Kearl, et al., each answer was given a score (strongly agree = 3, agree with provisions = 2, disagree = 1) and the mean score was calculated for each group. The means for each group on each question were compared using a t-test. When the mean scores were not significantly different at a 95% level the groups were assumed to be in agreement.

In the macroeconomic categories, economics students agreed with professionals on two of five positive propositions, and on only one of five normative items. Here the economics students looked much like their non-economics peers who had agreed with professionals on two of five positive and two of five normative propositions.

One result of particular interest was that economics students disagree with professional economists on all four items (3), (12), (16), (19) pertaining to money and monetary policy. Economic students' responses, however, unlike those of their non-economics peers, were systematic. They were consistently more "monetarist" in their responses than the professional economists reflecting, perhaps, the monetarist orientation of the instructor who teaches most of our introductory macro classes.

Since we did not assess student's responses prior to their taking economics, we cannot rule out the possibility that students who took economics were pre-disposed in favor of market mechanisms and monetarism. These results, however, suggest that there is an "economic way of thinking" and that agreement among economists is a result of their education. Lessons about the benefits of market allocation and the costs of interference seem to take hold and influence our level of agreement. Our results suggest these lessons are learned with only limited exposure to economics.

Steve Jackstadt
Lee Huskey
P. J. Hill
Associate Professors of Economics
University of Alaska
Anchorage, Alaska

Journal of Economic Perspectives—Volume 2, Number 4—Fall 1988—Pages 213–222

Notes

The annual meeting of the **American Economic Association** will be held in New York, NY, December 27–30, 1988.

The Professional Placement Service will be located at the New York Marriott Marquis Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; 9:00 A.M. to 12:00 noon, December 30.

Call for Papers for the 1989 AEA Meeting. Members wishing to give papers or make suggestions for the program for the meetings to be held in Atlanta, GA, December 27–30, 1989, are invited to send their ideas to Elton Hinshaw, American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. The sessions sponsored by the American Economic Association will be of two types: invited papers and contributed papers. A few of the invited papers will be published in the *Papers and Proceedings* issue to appear May 1990; the contributed papers will not be published. Suggestions of topics or proposals for papers for sessions should be submitted as soon as possible. Include the bibliographic code (*Journal of Economic Literature* codes) for each suggested topic, paper proposal and abstract. All proposals will be grouped by the *JEL* code and sent to the appropriate member of the Program Committee. Abstracts of proposed contributed papers must be received no later than February 1, 1989. Econometric studies or highly mathematical papers are not appropriate for the sessions sponsored by the AEA; members wishing to present such papers may submit their proposals or abstracts to the Econometric Society.

Call for papers. The **Southwestern Economics Association** will hold its annual meeting at the Excelsior Hotel in Little Rock, AR, March 29–April 1, 1989, in conjunction with the Sixty-seventh Annual Meeting of the Southwestern Social Science Association. The meeting's theme is "Vanishing Borders: America in the World Community." Those interested are invited to submit proposals for papers and panel discussions or to express their interest in serving as a chairperson or discussant. Send abstracts or other information before November 10, 1988, to: Luvonia J. Casperson, President Elect, Southwestern Economics Association, Department of Economics and Finance, Louisiana State University in Shreveport, One University Place, Shreveport, LA 71115.

Call for papers. The **Southern Economic Association** will hold its annual meeting at the Marriott's Orlando World Center, Orlando, FL, November 19–21, 1989. Deadline for submission of two copies of one-page paper abstracts and cover sheets (with $15 submission fee) is February 1, 1989. For details about subject categories and submitting an abstract or volunteering as chairperson or discussant, write: Dr. Joseph M. Jadlow, College of Business Administration, Oklahoma State University, Stillwater, OK 74078.

Call for papers. The **Economic and Business Historical Society** will hold its 14th annual meeting at the Radisson Francis Marion Hotel in Charleston, SC, April 27–29, 1989. The EBHS encourages interdisciplinary scholarship in economics and business history. Its international membership includes those with interests in economics, history, business, social science, geography, and other disciplines. Members and nonmembers are invited to submit papers, offer their services as session chairs and discussants. Please enclose a two-page abstract with proposals for papers. Write to Professor Jack Blicksilver, President and Program Chair, EBHS, Department of Economics, Georgia State University, University Plaza, Atlanta, GA 30303.

The Business Association of Latin American Studies will hold its annual meeting on February 15–18, 1988, in Boca Raton, Fl. For further information contact Dr. Robert P. Vichas, Conference Director, Florida Atlantic University, Business Association of Latin American Studies, P.O. Drawer 7638, Fort Lauderdale, FL 33338.

Call for papers. The ***International Journal of Manpower*** solicits papers in any area of manpower management, including applied microeconomic studies of the labor market. Articles should be submitted in duplicate to Dr. G. Johnes, Assistant Editor, IJM, Department of Economics, Gillow House, The University, Lancaster, LA1 4YX, United Kingdom.

Call for papers. ***Religion***, an international journal served by both European and American editorial boards, invites manuscripts on ideologies, religions, values and world views, continuing the tradition of inquiry into economics and values. It furthers conceptual and empirical investigations like gift exchanges, individualism, sacrifice, voluntary poverty, and world renunciation. Send manuscripts (3 copies, postage for returns) to Dr. Ivan Strenski, 3463 Meier Street, Mar Vista, Los Angeles, CA 90066-1701.

Call for papers. The **Law and Society Review** plans a special issue in early 1990 on Law and the Family. It is soliciting papers reporting socio-legal research on any aspect of legal intervention in the family. The *Review* will welcome submissions from the social sciences, including economics. All submissions (three copies are required) will be refereed. Inquiries should be sent to Professor Herbert Jacob, Department of Political Science, Northwestern University, Evanston, Il, 60208. Phone: 312-491-2648. Submission deadline is March 15, 1989.

Call for papers. The journal ***Population and Environment*** publishes papers ranging the broad, interdisciplinary perspective on the relationships between population and societal, cultural and physical environments. Demographic variables linked to life style, economics, law, health, business, and international relations are among those represented. Send inquiries and submissions to: Virginia Abernethy, Department of Psychiatry, School of Medicine, Vanderbilt University, Nashville, TN 37232.

The **National Association of Student Anthropologists** (NASA) was incorporated as a unit of the American Anthropological Association (AAA) in 1986 to address the needs and concerns of students of anthropology. NASA's Executive Committee and the general membership are comprised of representatives from all areas of anthropology (e.g., economic anthropology, social anthropology, medical anthropology, etc.) We want to expand our involvement with students of natural and social sciences who are invited to become members of both the AAA and NASA to demonstrate the interrelatedness of our disciplines within the framework of interdisciplinary scholarship. Write to: American Anthropological Association, 1703 New Hampshire Avenue, NW, Washington, DC 20009.

Call for papers. A conference on "The Worker in Transition: Technological Change" will be held in Washington, DC., April 5–April 7, 1989. Cosponsoring it is the Consortium of Social Science Associations, the American Society of Mechanical Engineers, the Labor Department (Bureau of Labor Management Relations) and Texas A & M University. The conference will be concerned with issues relating to changes in work and the workforce brought about by the rapid technological developments occurring around the world, with emphasis on the individual worker. The purpose of the conference is to bring together contributions from individuals in both technical and nontechnical disciplines which focus on the human implications of present and future changes in employment and the workforce. The program committee will accept papers relating to a wide range of topics, such as: accelerating rates of technological change, effects on engineers and managers as professionals, responses to major workforce reductions, public policy responsibilities and responsibilities of industry. The deadline is December 1, 1988. Paper will be peer-reviewed and authors will be notified of the acceptance of their papers by January 6, 1989. Papers accepted for the conference will be published in bound conference proceedings by the ASME. Manuscripts should be submitted to the program chairman: Professor Thomas J. Kozik, Institute for Innovation and Design in Engineering, Texas A & M University, College Station, TX 77843-3123, USA.

Call for papers. The **Association of Managerial Economists** will hold its seventh annual meeting in Atlanta, GA, December 28–30, 1989. Three sessions of competitively selected papers will be featured in conjunction with the Allied Social Science Associations (ASSA) Annual Meeting. Theoretical, empirical and policy analyses across a broad spectrum of topics will be featured. Especially encouraged are papers integrating the theory of accounting, finance and industrial organization in topic areas such as the market for corporate control (buyouts, mergers, share repurchases, spinoffs), managerial labor markets, measuring monopoly profits, and the information content of accounting data. Both members and nonmembers are invited to submit papers, offer their services as session chairs and discussants, and/or make program suggestions to: Professor Mark Hirschey, AME Program Chair, School of Business, University of Kansas, Lawrence, KS 66045, phone: 913-864-7538. Submission deadline is July 15, 1989.

Call for papers. The 1989 meeting of the **History of Economics Society** will be held at the University of Richmond, Richmond, VA, June 10–13. Proposals to deliver papers (including an abstract of 250 words), organize sessions, and/or act as discussant should be sent by December 1, 1988, to President-elect William J. Barber, Department of Economics, Wesleyan University, Middletown, CT 06547. A selection of papers presented at the 1989 Conference will be published in the series on *Perspectives on the History of Economic Thought*.

Call for papers. ***Advances in Econometrics***, a research annual, is soliciting abstracts for papers in the area of Econometric Methods and Models for Industrial Organization. Papers chosen for publication will appear in the 1990 volume *Advances in Econometrics: Econometric Methods and Models for Industrial Organization*. This volume will publish papers that extent economists' ability to use quantitative techniques in analysis of industrial organization issues, including: comparing effects of regulation and deregulation, creating and assessing antitrust policies, describing and measuring the impact of market share on market structure, and simulating outcomes of various policy proposals. Papers developing techniques as well as papers presenting theoretical and empirical work will be considered. Please send your proposed abstracts by March 15, 1989, to Professor George F. Rhodes, Jr., Department of Economics, Colorado State University, Fort Collins, CO 80523.

Call for papers. A special conference on "Health Economics and Health Policy in the 1990s: Surprises from the Past, Forecasts for the Future" will be held in Williamsburg, VA, April 13–14, 1989, by the Williamson Insttute, the Medical College of Virginia, University of California at Berkeley, Project HOPE, and SRI International. The conference will examine where health economics and health policy have been, where they are going, and reflect on past questions, surprise answers from health services research and the important issues for the next decade. For further information write: Louis F. Rossiter, Ph D., Associate Professor and Director, Williamson Institute for Heath Studies, Medical College of Virginia/VCU, P.O. Box 203-MCV Station, Richmond, VA 23298-0206.

Call for papers. The Department of Economics, **American University**, will hold a conference, entitled "Global Imbalances: Alternative Perspectives on the International Economy", May 27–29, 1989. The department encourages papers on global trade and financial imbalances, Third World impoverishment and adjustment policies, transnational corporations and the internationalization of capital, labor rights and development, NICs and export-led growth, contemporary macroeconomic management, militarization and development. Send proposals to: John Willoughby, Dept. of Economics, American University, Washington, DC 20016, USA, by *Dec. 31, 1988.* Contact Mr. Willoughby for more information.

Call for papers. A "Workshop on Economic Cycles and Growth: Stylized Facts and Explanatory Models" will be held at Lake Balaton, Hungary, August 24–26, 1989. The workshop will encourage empirically relevant work on economic dynamics at the macro and micro levels. For further information contact: Attila Chikan, ISIR, V. Veres Palne u. 36, Budapest H-1053, Hungary. Phone: 36-1-174-106.

Call for papers. The **Jerome Levy Economics Institute** of Bard College is sponsoring a conference on "Profitability, Stability and Measurement" at Bard College, March 16–18, 1989. Those interested in presenting papers at the conference should contact: Prof. Dimitri Papadimitriou, Jerome Levy Economics Institute of Bard College, Annandale, NY 12504. Phone: 914-758-2141.

Call for papers. The **European Economic Association** will hold its fourth annual congress in Augsburg, West Germany, September 2–4, 1989. Papers in all areas of economics, including the history of economic thought and economic history, are sought. Submissions should include two copies of the paper and a one-page abstract, and must be sent before March 1, 1989, to: Martin Hellwig, chairman, Program Committee, EEA Conference, University of Basel, Institute fur Volkswirtchft, CH 4003 Basel, Switzerland.

Call for papers. The **Risk Theory Seminar** is a forum for critical discussion of research issues in risk and insurance and encourages high quality research in those fields. As membership is obtained by presenting a research paper to the Seminar, the program committee will choose eight or nine papers for presentation at the 1989 Seminar to be held April 14–16 at the University of Utah, Salt Lake City, UT. Papers may be theoretical, empirical, or cover applications, with all areas of risk and insurance eligible as topics. Past papers have included economic, financial, actuarial and behavioral perspectives. Those interested in participating should send 3 copies of a detailed outline of their proposed presentation by *December 15, 1988, to* Dr. Stuart Klugman, Drake University, College of Business Administration, Des Moines, IA 50311.

The ***Review of Urban and Regional Development Studies*** is a new refereed journal published biannually by the Applied Regional Science Conference. The journal welcomes policy oriented papers dealing with urban and regional development issues with emphasis on developing countries. Send inquiries and submissions to: Professor Koichi Mera, Managing Editor, Tokyo International University, Matoba-kita, Kawagoe, Saitama-ken, 350 Japan.

The **Jerome Levy Economics Institute** of Bard College is accepting applications for one-year resident research fellowships to begin in June or September 1989. The institute is interested in research proposals concerning the macroeconomic determination of profits and other research dealing with issues of unemployment, inflation, growth and development, government spending and tax policies, capital investment and productivity, business cycles and economic stability, and other areas with important policy implications. For more information, contact Dimitri Papadimitriou, Executive Director, Jerome Levy Economics Institute of Bard College, Annandale, NY 12504.

The annual **Conference of the International Institute of Public Finance** (IIPF) will be held in Buenos Aires August 28–31, 1989. The subject field of the conference will be "Public Finance and Steady Economic Development". For more information contact the IIPF Executive Secretary, Mrs. B. Schneider, Saar University, D-6600 Saarbrücken 11, Federal Republic of Germany.

Pre- or postdoctoral graduate students are invited to submit papers for the 1989 **Dorothy S. Thomas Award** competition held by the Population Association of America. The award of $1,500 will be presented annually for the best paper in the field of internal migration or the interrelationships among social, economic, and demographic variables. For information contact Susan C. Watkins, Department of Sociology, Univerity of Pennsylvania, Philadelphia, PA 19104-6298. Papers must arrive by January 15, 1989.

The Economics Department and the School of Natural Resources of the **University of Michigan** offer a joint PhD Program in Natural Resource Economics. They also offer Masters degrees which provide an opportunity to pursue Natural Resource Economics. For further information, write: Natural Resource Economics Program, Economics Department, Lorch Hall, the University of Michigan, Ann Arbor, MI 48109.

The 1989 Annual Meeting of the **Urban Affairs Association** will be held in Baltimore, MD, March 15–18, 1989. The Association is a professional organization for those engaged in education, research or service in urban affairs. The conference theme is "Urbanization in the Nineties: The Closing of a Century" and will include, political social, economic, and demographic issues as well as the role and operation of government. For information, contact: Drs. James R. Bohland and Patricia Edwards, 1989 UAA Program Chairs, Urban Affairs and Planning, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061.

Fellowship Announcement. Research opportunities are available through the American Statistical Association (ASA)/National Science Foundation (NSF)/Census Research Program, where senior researchers and junior research associates conduct projects in residence at the Census Bureau. Fellowship candidates submit research proposals that are competitively evaluated by a program Review Board. New fellows may choose associates from a pool of applicants. Research can be done in any area related to Bureau data or methodology, such as development of public use files of economic microdata as well as various demographic and methodological studies. Fellowship terms are flexible; salaries are commensurate with qualifications. Application deadlines are early January 1989 for fellowships and mid-February for associateships. For more information, contact Dr. Robert C. McGuckin, Center for Economic Studies, Bureau of the Census, Room 3442-3, Washington, DC 20233. Phone: 301-763-2337.

The ***Review of Political Economy*** is a new journal designed to encourage contributions from non-orthodox groups such as Post-Keynesian, institutionalist, subjectivist, and behavioral economists. Submission from economists doing creative work in the broad traditions of political economy is also welcome. Every two years the journal will confer the G.L.S. Schackle prize for the best essay written by a young scholar on a subject of interest to journal readers. The *Review* will also award an Eichner Travel Scholarship every two years for a promising postgraduate student to present a paper at the *Review*'s annual conference. For more information contact John Pheby, Centre for Post Keynesian Studies, Department of Government and Economics, City of Birmingham Polytechnic, Birmingham B42 2SU, England.

The **Foreign Policy Research Institute** will hold its eighth annual competition for the Thornton D. Hooper Fellowship in International Security Affairs. The competition seeks to promote original research on issues affecting U.S. and international security matters. Economics, international relations specialists, political scientists and members of related disciplines are invited to apply. The selected applicant will spend a year in residence at the Foreign Policy Research Institute beginning Fall 1989, and receive a stipend. Applicants should submit copies of three publications, or writings accepted for publication, a curriculum vitae, three letters of recommendation and a six-page, double-spaced research proposal on some aspect of international security affairs. Fellows must have received or be about to receive a doctoral degree or have equivalent experience. Application deadline is December 15, 1988. Submit applications to: Hooper Fellowship Competition, Foreign Policy Research Institute, 3615 Chestnut Street, Philadelphia, PA 19104.

The **Missouri Valley Economic Association** will hold its twenty-fifth anniversary annual conference February 23–25, 1989, at the Allis Plaza Hotel, Kansas City, MO. Persons interested in being on the program should contact Wallace Peterson, Department of Economics, University of Nebraska-Lincoln, Lincoln, NE 68588. Persons wanting membership information should contact Pauline Fox, MVEA Secretary, Department of Economics, Southeast Missouri State University, Cape Girardeau, MO 63701.

Finance & Development, published jointly by the International Monetary Fund and the World Bank, is a quarterly for a diverse readership that consists of policymakers, academics, economic practitioners, and those interested in the work of the two institutions. It publishes articles of an informative or analytical character on the policies and activities of the institutions, emphasizing new aspects or approaches as they arise as well as general pieces on financial and developmental issues. The journal is available without charge to members of the AEA who request it. Please state that you are a member of the AEA and address your request to: Subscription Services, *Finance & Development*, International Monetary Fund, Washington, DC 20431. The journal is available in English, Arabic, Chinese, French, German, Portuguese and Spanish. Please specify the language edition you wish to receive.

The **Sydney Christian Economists Group** (SCEG) is a fellowship of evangelical Christians concerned with understanding economic matters in the light of Biblical theology and ethics. As a professional society, SCEG seeks to provide encouragement and support to Christians practicing in the economics profession. For information, contact Secretary, SCEG, Box 207, Wentworth Building, University of Sydney, NSW, Australia 2006.

Retirements

W. H. Locke Anderson, professor emeritus of economics, University of Michigan, Department of Economics, July 1, 1988.

Robert R. Dince, professor of banking and finance, University of Georgia, December 31, 1987.

Daniel Fusfeld, professor emeritus of economics, University of Michigan, Department of Economics, May 31, 1987.

John W. Kendrick: professor of economics, George Washington University, July 1, 1988.

James Morgan, professor emeritus of economics, University of Michigan, Department of Economics, December 31, 1987.

Joel Sailors, emeritus, Department of Economics, University of Houston, April 27, 1988.

Foreign Scholars

Pulapre Balakrishnan, University of Oxford: visiting assistant professor, University of Michigan, Department of Economics, September 1, 1988.

Katalin Ferber: visiting lecturer, Wilfrid Laurier University, September 1, 1988. Currently working as a researcher at the Historical Institute in Hungary.

Nadav Halevi, Department of Economics, The Hebrew University of Jerusalem: visiting professor in economics, Department of Economics, Wesleyan University.

Yuzuru Ima, Kwansei Gakuin University, School of Business Administration, Japan: visiting professor, Department of Economics and Business, North Carolina State University, March 29, 1988–March 28, 1989.

Peter Morgan, University of Western Ontario: visiting associate professor of economics, University of Michigan, College of Literature, Science and the Arts, September 1, 1988.

Prasanta K. Pattanaik, University of Birmingham: visiting professor of economics, The University of Albany, State University of New York, September 15, 1988–January 15, 1989.

Alasdair Smith, University of Sussex: visiting professor of economics, University of Michigan, College of Literature, Science and the Arts, September 1, 1988.

Promotions

Berhanu Abegaz: associate professor, economics department, College of William and Mary, August 1988.

W. James Adams: professor of economics, University of Michigan, College of Literature, Science and the Arts, September 1, 1988.

Michael R. Baye: associate professor, Department of Economics, Texas A & M University, September 1, 1988.

Michael T. Belongia: research officer, Federal Reserve Bank of St. Louis, January 1, 1988.

Donald S. Elliott: professor, Department of Economics, Southern Illinois University at Edwardsville, July 1987.

Bryan L. Boulier: professor of economics, George Washington University, July 1, 1988.

Cheryl A. Casper: professor of economics, Kent State University, August 21, 1988.

Kenneth S. Chan: professor of economics, McMaster University, July 1, 1988.

Fwu-Ranq Chang: associate professor of economics, Indiana University Bloomington, July 1, 1988.

Steven Craig: associate professor, Department of Economics, University of Houston, August 29, 1988.

Vijaya Duggal: tenure as an associate professor of economics, Widener University, July 1988.

David H. Feeny: professor of economics, McMaster University, July 1, 1988.

Alexander J. Field: professor, Department of Economics, Santa Clara University, September 1, 1988.

David W. Gillen: professor of economics, Wilfrid Laurier University, July 1, 1988.

Eitan Gerstner: associate professor, Economics and Business, North Carolina State, August 16, 1988.

William J. Hausman: professor, economics department, College of William and Mary, August 1988.

Charles E. Hegji: associate professor, Department of Economics, Auburn University at Montgomery, Montgomery, AL, September 1, 1988.

James Hess: professor, Economics and Business, North Carolina State University, August 16, 1988.

Eric R. Jensen: associate professor, economics department, College of William and Mary, August 1988.

John Komlos: associate professor of history and economics, University of Pittsburgh, April 1988.

John Laitner: professor of economics, University of Michigan, College of Literature, Science and the Arts, May 31, 1987.

Jack E. Leach: associate professor of economics, McMaster University, July 1, 1988.

Karen Leppel: associate professor of economics, Widener University, July 1988.

Christopher Lingle: senior lecturer, Department of Economics, University of Natal, Durban, South Africa, February 1, 1988.

Lonnie W. Magee: associate professor of economics, McMaster University, July 1, 1988.

John J. Maher: Institutional Analyst, Union Theological Seminary, July 1, 1987.

James C. Murdoch: associate professor, Department of Economics, Auburn University at Montgomery, Montgomery, AL, September 1, 1988.

Alireza Nasseh: associate professor, Department of Economics, Saint Louis University, July 1, 1988.

Douglas K. Pearce: professor, Economics and Business, North Carolina State University, August 16, 1988.

Carl P. Simon: professor of economics and mathematics, University of Michigan, College of Literature, Science and the Arts, September 1, 1988.

Stephen C. Smith: associate professor of economics, George Washington University, July 1, 1988.

William B. Walstad: professor of economics, University of Nebraska-Lincoln, May 1988.

Stuart E. Weiner: research officer and economist, Economic Research Department, Federal Reserve Bank of Kansas City, January 1, 1988.

Donald R. Williams: associate professor of economics, Kent State University, August 21, 1988.

Larry N. Willmore: head, Industrial Development Unit of the Mexico office, United Nations Economic Commission for Latin America and the Caribbean, January 1988.

Michael K. Wohlgenant: professor, Economics and Business, North Carolina State University, July 1, 1988.

Martin Zimmerman: professor of business economics and economics, University of Michigan School of Business Administration and College of Literature, Science and the Arts, September 1, 1988.

Administrative Appointments

Virginia Abernethy: Editor, *Population and Environment*, Fall 1988.

Bala Batavia: acting dean, professor, College of Commerce, Depaul University, July 1, 1988.

Bryan L. Boulier: chairman, Department of Economics, George Washington University, July 1, 1988.

Charles Brown. associate chair, Department of Economics, College of Literature, Science and the Arts, University of Michigan, September 1, 1987.

Than Van Cao: chairman, Department of Management and Economics. School of Business and Economics, Eastern Montana College, September 1988.

Allin F. Cottrelli: chair, Department of Economics, Elon College, June 1, 1988.

Joseph DiAngelo: dean of the School of Management, Widener University, January 1988.

William L Marr: assistant dean (Graduate Studies)/associate director (Research and Instructional Development), Wilfrid Laurier University. May 1, 1988.

Charles F. Revier, associate professor, Department of Economics, Colorado State University; associate professor and chair, Department of Economics, Colorado State University, July 1, 1988.

A Leslie Robb: associate chairman, Department of Economics, McMaster University, July 1, 1988.

Thomas K. Tiemann: dean, Martha and Spencer Love School of Business, Elon College, January 1, 1988.

Charle Waldauer: head of the Economics Department, Widener University, July 1988.

Carl E. Walsh: chairman, Economics Department, University of California, Santa Cruz, July 1, 1988.

Jack W. Wilson: department head, University Studies, North Carolina State University, July 1, 1988.

New Appointments

Bruce Arnold, State Department and Princeton University: economist, National Security and International Affairs Division, General Accounting Office, May 1988.

R. Dale Ballou, Yale University: assistant professor, Department of Economics and Business, North Carolina State University, August 16, 1988.

John C. Beghin, University of California-Berkeley: visiting assistant professor, Department of Economics and Business, North Carolina State University, June 1, 1988.

Avner Ben-Ner, W. Averell Harriman School for Management and Policy, State University of New York at Stony Brook: associate professor, Industrial Relations Center, Carlson School of Management, University of Minnesota, September 16, 1988,

Kenneth Binmore, London School of Economics and Political Science: professor of economics, University of Michigan, College of Literature, Science and the Arts, September 1, 1988.

David S. Bizer: instructor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Carol E. Bray, First Interstate Bancorp: economist, Resources, Community and Economic Development Division, General Accounting Office, Summer 1988.

Jaeho Cho, University of Pennsylvania: assistant professor, Baruch College CUNY, August 1988.

David L. Cleeton: associate professor and chairman, Department of Economics, Oberlin College, Fall 1988.

Bruce P. Corrie: assistant professor of economics, Department of Business and Economics, Concordia College.

Ben Craig, assistant professor of economics, Washington State University: assistant professor of economics, Indiana University-Bloomington, August 1988.

David J. DeOrio: financial economist, U.S. Department of the Treasury, Office of Tax Analysis, March 1988.

Enrica Detragiache: instructor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Rose Anne Devlin: lecturer, Wilfrid Laurier University, July 1, 1988.

Robert R. Dince: visiting scholar, Center for Banking, East Tennessee State University.

James J. Dowling, MIT: assistant professor, Baruch College CUNY, August 1988.

Martine Duchatelet, Fordham University: visiting associate professor, Baruch College CUNY, August 1988.

Larry Epstein: visiting professor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Christina A. Fader: lecturer, Wilfrid Laurier University, September 1, 1988.

Philip Farah, University of New Mexico: economist, Resources, Community and Economic Development Division, General Accounting Office, Summer 1988.

Emil Friberg, University of North Carolina: economist, National Security and International Affairs Division, General Accounting Office, September 1988.

J. A. Galbraith, Research Branch, Library of Parliament, Ottawa, Canada: adjunct research professor, Department of Economics, Carleton University, Ottawa, July 1, 1988.

David Harrison: senior consultant, National Economic Research Associates, Cambridge, MA, February 16, 1988.

Anne Hill, Rutgers University: visiting assistant professor, Baruch College CUNY, August 1988.

Mary Claire Huber: lecturer, University of Michigan, Department of Economics, September 1, 1988.

Muhammed N. Islam: visiting associate professor, Wilfrid Laurier University, July 1, 1988.

Stephen Jones, University of British Columbia: visiting associate professor of economics, McMaster University, July 1, 1988.

David Joulfain: financial economist, U.S. Department of the Treasury, Office of Tax Analysis, Fall 1987.

Frederick L. Joutz, Oberlin College: assistant professor, Department of Economics, George Washington University, September 1, 1988.

John W. Keating, Northwestern University: assistant professor, Department of Economics, George Washington University, September 1, 1988

M. Ali Khan: professor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Jun-Il Kim: assistant professor of economics, University of California, Santa Cruz, July 1, 1988.

Thomas J. Kniesner, associate professor of economics, University of North Carolina at Chapel Hill: professor of economics, Indian University-Bloomington, August 1988.

Elaine Koppana: visiting assistant professor, economics department, College of William and Mary, August 1988–May 1989.

Richard Krashevski, AFL–CIO: economist, Office of the Chief Economist, General Accounting Office, April 1988.

Gene Kuehneman, Syllogistics, Inc.: economist, Office of the Chief Economist, General Accounting Office, April 1988.

Peter Kuhn, University of Western Ontario: associate professor of economics, McMaster University, July 1, 1988.

James P. Lesage: assistant professor, Department of Economics, University of Toledo, Fall 1988.

Brian Lewis: lecturer, Wilfrid Laurier University, July 1, 1988.

Eng Seng Loh, Wittenberg University: assistant professor, Department of Economics, Kent State University, August 21, 1988.

Willard Manning, The Rand Corporation: professor of Economics and Public Health, University of Michigan, College of Literature, Science and the Arts and the School of Public Health, September 1, 1988.

John G. Marcis: assistant professor, Department of Economics, University of Toledo, Fall 1988.

Xavier Maret: lecturer, University of Michigan, Department of Economics, September 1, 1988.

Paul D. McCann, University of Pennsylvania: assistant professor, Baruch College CUNY, August 1988.

Walter S. McManus, University of Florida at Gainesville: associate professor, Baruch College CUNY, August 1988.

Thomas Melito, Columbia University: economist, National Security and International Affairs Division, General Accounting Office, September 1988.

Yesook Merrill, SUNY-Binghamton: economist, Office of the Chief Economist, General Accounting Office, April 1988.

Michael O. Moore, University of Wisconsin-Madison: assistant professor, Department of Economics and Lloyd Elliot School of International Affairs, George Washington University, September 1, 1988.

Maria Muniagurria: assistant professor of economics, University of California, Santa Cruz, July 1, 1988.

Sandra Peart: instructor, economics department, College of William and Mary, August 1988.

Rodrigo Peruga, University of California, San Diego: assistant professor of economics, Indiana University-Bloomington, August 1988.

Robert Petersen: senior analyst, National Economic Research Associates, White Plains, NY, October 26, 1987.

Irene Powell, Wellesly College: economist, Human Resources Division, General Accounting Office, Summer 1988.

Kalman Rupp, Abt Associates: economist, Human Resources Division, General Accounting Office, June 1988.

Jolyne (Melmed) Sanjak: assistant professor, Department of Economics, The University at Albany, State University of New York, September 1988.

Carlos P. Santiago: associate professor, Department of Economics, The University of Albany, State University of New York, September 1988.

Gary J. Santoni, Federal Reserve Bank of St. Louis: professor of economics, Ball State University, August 1988.

Chera Sayers: assistant professor, Department of Economics, University of Houston, August 29, 1988.

Linda Stanley, Dept. of Ag. & Resource Economics, Colorado State University: assistant professor, Department of Economics, Colorado State University, August 20, 1988.

Anne Stevens, University of Maryland: economist, General Government Division, General Accounting Office, September 1988.

Steven M. Suranovic, Cornell University: assistant professor, Department of Economics, George Washington University, September 1, 1988.

Joseph Swierzbinski, University of Washington: assistant professor of economics and resource/forest economics, University of Michigan, College of Literature, Science and the Arts and School of Natural Resources, September 1, 1988.

Y. Edwin Tang, The University of Texas at Dallas: visiting assistant professor, Department of Economics and Business, North Carolina State University, August 16, 1988.

Evan Tanner: visiting assistant professor, Department of Economics, Oberlin College, Fall 1988 semester only.

Doki K. Tano: assistant professor, Department of Economics, University of Toledo, Fall 1988.

William Taylor, Bell Communications Research: vice president, National Economic Research Associates, February 29, 1988.

Robert Thomas: lecturer, University of Michigan, School of Business Administration and College of Literature, Science and the Arts, September 1, 1988.

Velma Montoya Thompson: associate professor of finance, Faculty of Finance, Real Estate and Law, California State University Polytechnic, Pomona, September 1988.

Randall M. Wigle: assistant professor, Wilfrid Laurier University, July 1, 1988.

Martin Zelder: instructor, economics department, College of William and Mary, August 1988.

Leaves for Special Appointments

James Hess, North Carolina State University: visiting professor, German Research Association Chair of Organized Theory, Koblenz School of Corporate Management, Koblenz, Germany, August 16, 1988–May 15, 1989.

David N. Hyman, North Carolina State University: guest scholar, Brookings Institution, Washington, D.C., August 16, 1988–December 31, 1989.

Dennis W. Jansen: Visiting Scholar at The Federal Reserve Bank of St. Louis, September 1, 1988–September 1, 1989.

Resignations

James R. Barth, Department of Economics, George Washington University; Federal Home Loan Bank Board.

Simón Teitel, The Catholic University of America, January 11, 1988.

Robert K. Triest: assistant professor, Department of Economics, The Johns Hopkins University, July 1, 1988.

Miscellaneous

Zvi Griliches, Professor of Economics at Harvard University, is the winner of the W. S. Woytinski Lectureship Award for 1989. This award, given in recognition of significant research contributions in economics, commemorates W. S. Woytinski whose career was devoted to empirical research on human resource policy issues.

Professor Baldev Raj will visit the Institute of Economics, Norwegian School of Economics and Business Administration, Bergen, Norway, June 15, 1988, to August 31, 1988, funded by a fellowship from the Norwegian Research Council for Science and the Humanities.

To Departmental Representatives and Executive Officers

When submitting information for inclusion in "Notes," please observe the following guidelines:

Calls for papers, notice of professional meetings, and other announcements of interest to economists should be submitted in one paragraph that contains all relevant information.

News of individual members should be labeled as to category: (1) deaths; (2) retirements; (3) foreign scholars visiting the US or Canada; (4) promotions; (5) administrative appointments; (6) new appointments; (7) leaves for special appointments; (8) resignations; (9) miscellaneous. Give individual's name, present place of employment, new title, new institution, and date on which change is to occur.

Deadlines for each issue are: Spring, December 15; Summer, March 15; Fall, June 15; Winter, September 15.

Please send all information to the *Journal of Economic Perspectives*, Department of Economics, Stanford University, Stanford, CA 94305-6072. We reserve the right to edit material received.

Accession Number.
......121814......
Date 7-12-89

AMERICAN ECONOMIC ASSOCIATION

1989 ANNUAL MEMBERSHIP RATES

Membership includes:

– a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

● Regular members with annual incomes of $30,000 or less	$42.00
● Regular members with annual incomes above $30,000 but no more than $40,000	$50.40
● Regular members with annual incomes above $40,000	$58.80
● Junior members (available to registered students for three years only) Student status must be certified by your major professor or school registrar	$21 00
● In Countries other than the U S A , Add $16 00 to cover postage.	
● Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association)	$8.40

Please enter my membership for the following period:

☐ **Jan.-Dec.** ☐ **April-March** ☐ **July-June** ☐ **Oct.-Sept.**

First Name and Initial ____ **Last Name** ____ **Suffix** ____

Address Line 1 ____

Address Line 2 ____

City ____

State or Country ____ **Zip/Postal Code** ____

MAJOR FIELDS (TWO ONLY)

LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*. "Classification System for Books."

☐ ☐

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____

____

Authorized Signature ____

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

American Economic Association

Summer Minority Program

at Temple University

June 2, 1989–July 28, 1989

A New Economics Program with Exciting Features

An Attractive Stipend–plus Books,
Room, Board, Tuition and Travel are
Available to Qualified Students.

For More Information Contact:

DEPARTMENT OF ECONOMICS
RITTER HALL ANNEX
TEMPLE UNIVERSITY
PHILADELPHIA, PA 19122

(215-787-8887)